AD NUMBER

AD027933

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; DEC 1953. Other requests shall be referred to Wright Air Development Center, Attn: ARDC, Wright-Patterson AFB, OH 45433.

AUTHORITY

AFAL ltr, 17 Aug 1979

WADC TECHNICAL REPORT 54-9

THE POWER OF STATISTICAL TESTS

E. S. KEEPING

UNIVERSITY OF ALBERTA

DECEMBER 1953

WRIGHT AIR DEVELOPMENT CENTER

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

WADC TECHNICAL REPORT 54-9

THE POWER OF STATISTICAL TESTS

E. S. Keeping

University of Alberta

December 1953

Aeronautical Research Laboratory
Contract No. AF33(616)-321
RDO No. 475-415

Wright Air Development Center
Air Research and Development Command
United States Air Force
Wright-Patterson Air Force Base, Ohio

Carpenter Litho & Prtg. Co., Springfield, O.
500 - 16 November 1954

FOREWORD

This report was prepared by Professor E.S. Keeping at the University of Alberta, Canada, on Contract No. AF33(616)-321 with the United States Air Force.

The work was done under Research and Development Order No. R475-415, and was initiated by Dr. Paul R. Rider, Chief Statistician, Aeronautical Research Laboratory, Wright-Patterson Air Force Base, Ohio.

ABSTRACT

The concept of power for statistical tests of hypotheses, and various ideas connected with it, are described and illustrated. The power is given for a number of the common statistical tests, and tables are supplied which facilitate decisions on the sample sizes necessary for detecting differences between means, variances, proportions defective, etc. with prescribed power.

PUBLICATION REVIEW

This report has been reviewed and is approved.

FOR THE COMMANDER:

Leslie B. Williams

LESLIE B. WILLIAMS
Colonel, USAF
Chief, Aeronautical Research Laboratory
Directorate of Research

TABLE OF CONTENTS

Chapter I	Preliminary Ideas and Definitions	1
Chapter II	Testing the Mean of a Sample from a Normal Population of Known Variance	16
Chapter III	Test the Variance of a Sample from a Normal Population	34
Chapter IV	Student's t-test	45
Chapter V	Tests for the Proportion Defective	58
Chapter VI	The F-test	69
Chapter VII	Distribution-free Tests	78

List of Tables

Table I	Size of Sample necessary to detect with Probability P a One-sided Difference in the Mean equal to $z\sigma$ (Normal Population).	20
Table II	Size of Sample necessary to detect with Probability P a Difference either way in the Mean equal to $z\sigma$ (Normal Population).	25
Table III	Size of Sample necessary for Probability P of not finding a Difference of $z\sigma$ in the Mean where none actually exists. (Normal Population).	33
Table IV	Size of Sample necessary to detect with Probability P a Variance Ratio k different from 1.	39
Table V	Power of the t-test for distinguishing between the Means of two Samples of Size N, with common σ.	54
Table VI	Power of Test for Proportion Defective.	63
Table VII	Values of the Standard Deviation Ratio detectable with Power P, with Samples of Size N.	71
Table VIII	Ratio of Standard Deviation of Lot Means to Standard Deviation of Population, detectable with Power 0.5 at Significance Level 0.05.	77
Table IX	Critical Values of $\Sigma\xi_i$ for Van der Waerden's Test	80
Table X	Probabilities of Error of the First Kind in using the Critical Values of Table IX.	83

Table XI	Critical Values $\bar{U}$ for the Mann and Whitney Test.	87
Table XII	Minimum N and Maximum r for testing the Proportion of Signs in Paired Differences.	89
List of Graphs		
Fig. 1	Power Function of an Ideal Test	6
Fig. 2	Typical Power Functions of a Real Test, Corresponding to different Regions of Rejection	7
Fig. 3	Two Rectangular Distributions	13
Fig. 4	Power Curves for One-Sided Normal Test	18
Fig. 5	Power Curves for Two-sided Normal Test	22
Fig. 6	Power Curves for Non-central Chi-square Test	27
Fig. 7	No title	29
Fig. 8	No title	31
Fig. 9	Power Curves for Chi-square Test of Variance	35
Fig. 10	Distributions of X_i under hypotheses H_o, H_1, H_o	40
Fig. 11	Power Curves for One-tailed t-Test,	49
Fig. 12	Power Curves for Test of Proportion Defective ($\pi_o = 0.5$)	60
Fig. 13	Power Curves for Test of Proportion Defective ($\pi_o = 0.2$)	62

1 PRELIMINARY IDEAS AND DEFINITIONS.

1.1 Statistical Estimation.

In many practical problems we examine a sample in order to be able to say something about the population from which the sample was taken. For instance, we may want to find out whether the mean breaking strength of a large consignment of steel rods is greater than a specified value, or whether a machine is turning out an unduly large proportion of pistons with diameters outside the specified tolerance limits. From a sample, and especially from a small sample, we cannot expect to get exact information about the population. All that we can hope to do is to determine the probability that a statement which we make about the population is true. Naturally we want this probability, other things being equal, to be as large as possible.

The types of problems which we can investigate statistically by sampling, fall into two main classes: problems of estimation and problems of the testing of hypotheses. In problems of estimation we are concerned with the numerical value of some characteristic of the population such as the mean breaking strength (for a population of rods), and we form the best estimate we can of this quantity from measurements on a random sample. The size of the sample and the particular statistic (a function of the measurements) which we use for estimation are more or less within our control. As regards size, the larger the sample, of course, the more accurate the estimate, but questions of time and expense often seriously restrict the size of a practicable sample, and in routine work samples as small as four or five are quite common. The statistic, or estimator as it is sometimes called, should possess certain desirable properties; and in particular should be consistent, unbiased and as efficient as possible. It is said to be consistent if its value tends, as the sample size increases indefinitely, to the true value of the characteristic which is being estimated. It is unbiased if its expected value (that is,

the arithmetic mean of its values for a very large number of similar random samples, all of the same size) is equal to the true value. The efficiency is measured by the variance of the values of the statistic for random samples of the same size; the less this variance the more efficient the statistic, since the standard deviation (the square root of the variance) is a measure of the order of magnitude of the error which may be expected when the unknown population value is estimated from the known sample statistic.

If several statistics are available for estimating the same population characteristic, we shall naturally choose the most efficient one, unless it is so much more troublesome or time-consuming to calculate that the gain in efficiency is more than offset by the loss in speed. Both the arithmetic mean and the median of a sample from a normal population are consistent and unbiased statistics for estimating the mean of the population, but the arithmetic mean is more efficient than the median. The latter, on the other hand, is somewhat easier to calculate. Again, the variance s^2 of a sample of size N from a normal population is, when multiplied by $N/(N - 1)$, an unbiased estimate of the population variance σ^2. Another estimate is provided by the range of the sample (the difference between the greatest and the least values in the sample), the estimated variance being equal to the square of the range multiplied by a factor which depends on the sample size and which is obtainable from tables for sizes up to 20. For small sample sizes the range is nearly as efficient as the sample variance and is much easier to calculate.

1.2 Confidence Intervals.

When estimating a characteristic of the population from a sample statistic it is very desirable to know how much trust we can place in the estimate. For many practically important kinds of estimate it is possible to calculate a confidence interval, within which, with a certain degree of confidence, we can claim that the true value will lie. If, on the basis of a sample, we calculate upper and lower 95% confidence limits a and b,

for the estimation of a parameter θ, we imply that the probability of the truth of the statement $b < \theta < a$ is 0.95. This probability is to be interpreted as referring to the relative frequency of correct statements among a very large number of such confidence statements, each made on the basis of a separate random sample of the same size. Each sample will give rise to its own confidence interval, which may or may not actually include the true value, but it will do so in 95% of the samples. We therefore stand only a 5% chance of being wrong in making the statement on the basis of a single sample.

1.3 Tests of Hypotheses.

Instead of trying to estimate the precise value of some population characteristic or parameter θ, we may be more interested in whether or not it exceeds or falls below a certain specified level, or whether it lies between definitely specified limits. We may, for example, want to know whether the mean breaking strength of a certain type of thread is at least 100 lb. wt., or whether in a large lot of machined parts at least 98% will have diameters within, say, 5 thousandths of an inch of a given value. In such cases we use random samples to test a certain hypothesis, generally called the null hypothesis, H_o. For the example of the thread, the null hypothesis might be that that true mean breaking strength $\mu \geq 100$ lb. wt. The alternative hypothesis H_1 might then be that $\mu < 100$ lb. wt., and we try to decide between these two hypotheses by making measurements on one or more samples. In discussing a statistical test it is well to be quite clear at the outset about the nature of the null hypothesis and of the alternative hypothesis.

On the basis of the sample measurements we have three possible courses of action. We can (1) reject the null hypothesis, (2) accept it, or (3) hedge, and say that the results are indecisive and that further samples should be taken. If, for any reason, we are limited to one sample of a fixed size, we are bound to adopt one of the first two courses. In two-sample tests, and sequential tests, further sampling is possible, but most of the classical statistical tests are based on the fixed-sample concept, and it is with this procedure that we shall be mainly concerned.

1.4 Errors of the First and Second Kind.

If we are limited to acceptance or rejection of the null hypothesis we can obviously go wrong in two ways, either by rejecting the null hypothesis when it is really true (this is called an error of the first kind) or by accepting the null hypothesis when the alternative hypothesis is true (this is called an error of the second kind). It is in practice always possible to devise the test so that the probability of an error of the first kind is definitely less than some fixed value less than 1 (it may be, for instance, less than 0.05). At the same time, we should like the probability of an error of the second kind to be as small as possible, and we try to devise the test accordingly. One test is said to be more powerful than another of the same size if it gives a greater probability of rejecting the null hypothesis when it is false, or, what comes to the same thing, a smaller probability of making an error of the second kind.

1.5 Assumption of Normality.

A test usually consists in calculating a certain statistic T from the observed sample values and observing whether or not T lies in a pre-determined region of values which is called the region of rejection. If it does lie in this region the null hypothesis is rejected. The region of rejection is determined from the unknown distribution of T when the null hypothesis is true, and for most tests in common use it is necessary, in order to specify the region of rejection, to assume that the measured variable is normally distributed in the parent population. There are some tests which do not require this or any other assumption about the population distribution and which are therefore known as distribution-free tests. Also, some work has been done on sampling from a rectangular population and from a skew distribution known as Pearson's Type 111, but for the most part the assumption of a normal population is regularly made. From a good deal of experimental evidence it appears that a moderate degree of departure from normality will not seriously affect the ordinary tests. If the departure is marked, it is often possible by transforming the variable (for example, by using log x in-

stead of x, or by using Fisher's transformation $z = 1/2 \log \{(1 + r)/(1 - r)\}$ instead of r) to make the new variable much more nearly normal in distribution.

1.6 The Power of a Test

A null hypothesis regarding the value of a parameter θ is said to be simple if it specifies the population completely. Otherwise, it is said to be composite. If a variable x is normally distributed in the population with known variance σ^2 but with unknown mean μ, the hypothesis that $\mu = \mu_0$ is a simple hypothesis, since the population distribution is then completely determined. The alternative hypothesis is that $\mu = \mu_1$ where μ_1 is different from μ_0. If we are prepared to consider any possible alternative, we must allow either $\mu_1 < \mu_0$ or $\mu_1 > \mu_0$, and this is called a two-sided alternative. Sometimes, however, we are interested only in the possibility that $\mu_1 > \mu_0$. We may, for example, want to know whether a new treatment or process will improve the quality of a certain material and we feel sure that the treatment cannot make it worse. In such cases, we have a one-sided alternative.

Let us suppose that we want to test the simple hypothesis $\theta = \theta_0$ against the simple alternative hypothesis $\theta = \theta_1$, using the statistic T. Let us also suppose, for convenience, that T is distributed continuously in the population of all possible samples of the given size N with a density function* $f(t \mid \theta_0)$ for $\theta = \theta_0$. If the rejection region is denoted by R, the size of the test is given by

$$\int_{(R)} f(t \mid \theta_0)\, dt = \alpha \ . \tag{1.1}$$

The power of the test is

$$P(\theta) = \int_{(R)} f(t \mid \theta) dt \tag{1.2}$$

for $\theta \neq \theta_0$. When $\theta = \theta_1$,

$$P(\theta_1) = 1 - \beta \ , \tag{1.3}$$

where β is the probability of an error of the second kind.

* - That is, the probability that T lies between t and t + dt for the given value θ_0 is $f(t \mid \theta_0)\, dt$. The vertical stroke may be read as "given".

If the distribution function of the statistic is $F(t \mid \theta_o)$, which is the probability of a value equal to or less than t on the hypothesis that $\theta = \theta_o$, we can write $f(t \mid \theta_o)\,dt = dF(t \mid \theta_o)$, and this notation applies even when the statistic takes only discrete values. The distribution function is then to be interpreted as meaning the sum of the probabilities for all discrete values of T equal to or less than t.

The function $P(\theta)$ is called the power function of the test. The ideal test would be one in which $\alpha = 0$ and $P(\theta) = 1$ for all $\theta \neq \theta_o$. The curve of $P(\theta)$ would resemble that in Figure 1, and there would be no errors of the first or second kind. The null hypothesis would never be rejected when true and always rejected when false. This happy state of affairs seldom, if ever, arises.

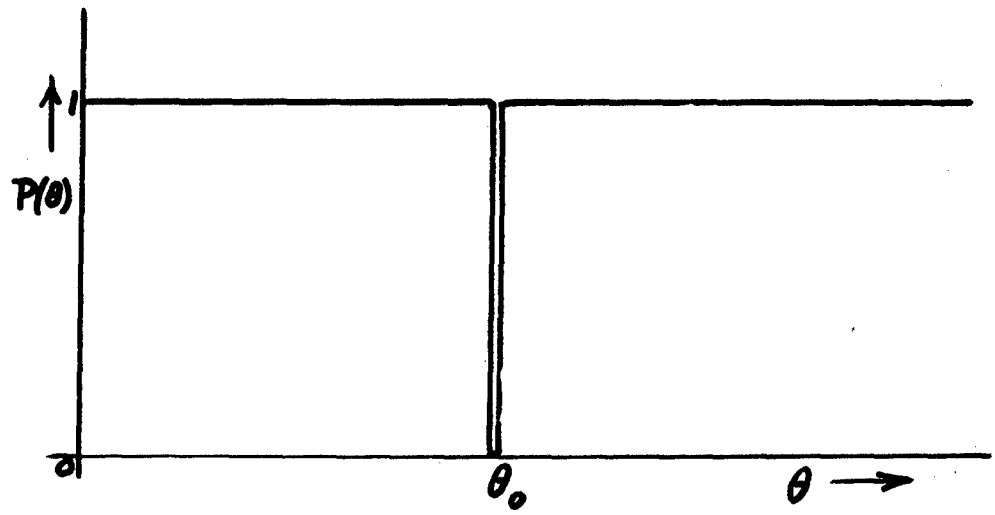


Fig. 1. Power Function of an Ideal Test.

Instead of $P(\theta)$, the function $\beta(\theta) = 1 - P(\theta)$ is often used. In this form it is called the operating characteristic (O.C., for short) of the test.

In practice, the power function of the test is more likely to resemble the curves of Fig. 2. If θ_1 is near θ_o, the power is small, which means that there is small probability of rejecting the hypothesis $\theta = \theta_o$ when θ is really equal to θ_1. However, in such a case no great harm is done, because we are only replacing the true value by a nearby one. When the distance between θ_o and θ_1 is large, a useful test will have a value of $P(\theta_1)$ near to 1.

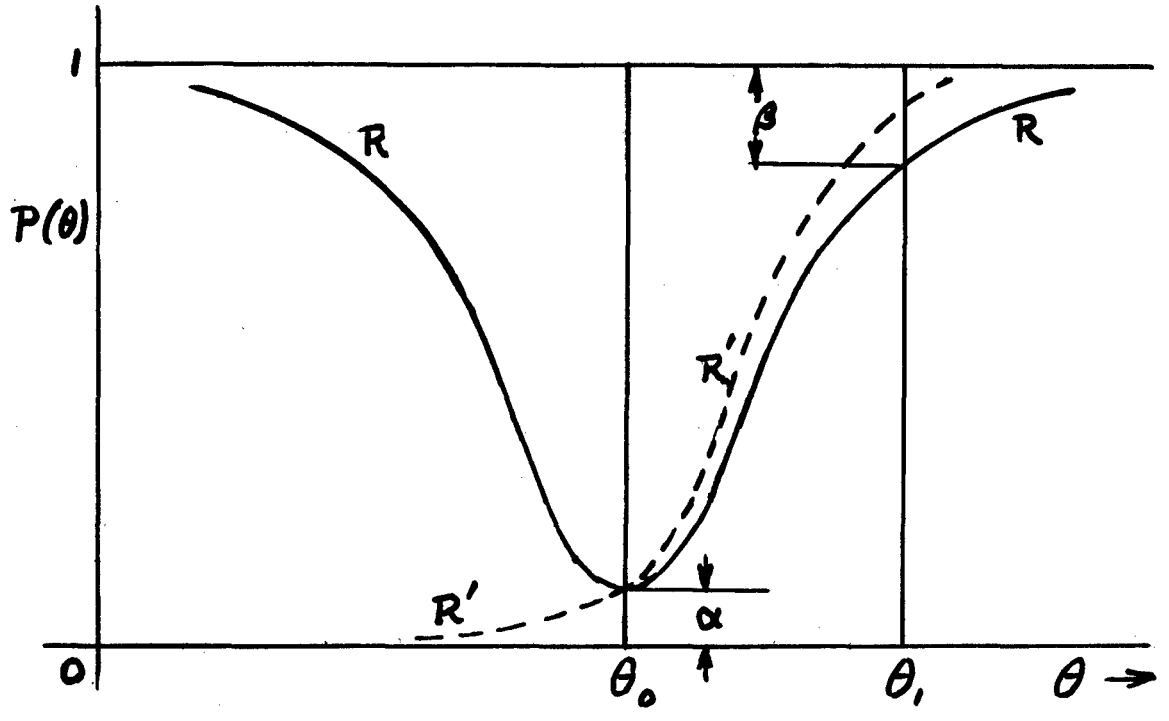


Fig. 2. Typical Power Functions of a Real Test, corresponding to different Regions of Rejection.

In general, the region of rejection R, for a given size α, can be chosen in many ways, each of which gives rise to a test with a different power function. If it happens that the test corresponding to a different region R' has a power curve which lies below that corresponding to R for all values of θ except θ_0, then the test using R is uniformly more powerful than that using R'. The probability α of an error of the first kind is the same for both, but the probability β of an error of the second kind is always less for R than it is for R'. That is, we shall in the long run more frequently go wrong if we use the test based on R' to distinguish between θ_0 and θ_1 than if we use the test based on R.

If this is true for every possible choice of R', then R is said to be a uniformly most powerful test, and if such a test can be found we shall be perfectly satisfied. Unfortunately, however, such tests are

seldom available, and what often happens is that R' will give a power curve which is above that of R for some values of θ and below it for other values, as illustrated in Fig. 2.

1.7 The Neyman - Pearson Lemma

We suppose that X is a random variable which can take values x lying in a certain region D. If X is discrete, like the number of spots on the upper face of a die, the possible values of x are isolated real numbers (e.g. 1, 2, 3, 4, 5, 6). If X is continuous, like a measured height, the possible values of x may form an interval or set of intervals of the real axis. If X is a set of N observations or measurements, the region D is N-dimensional. If R is a sub-region contained within D, the probability that x belongs to R is given by

$$Pr\ (x \in R) = \int_{(R)} dF(x \mid \theta) \tag{1.4}$$

where $F(x \mid \theta)$ is the cumulative distribution function for x and depends upon the value of a parameter θ (or possibly on several parameters).

If the variable X is continuous

$$dF(x \mid \theta) = f(x \mid \theta)\, dx$$

where $f(x \mid \theta)dx$ is the probability that x lies between x and x + dx for the given value of θ. If X is discrete $\int_{(R)} dF(x \mid \theta)$ is the sum of of the probabilities $f(x \mid \theta)$ for each of the distinct possible values of x which lie within the region R, for the given value of θ.

Suppose we now want to test the simple null hypothesis $\theta = \theta_0$, against the simple alternative hypothesis $\theta = \theta_1$. (It is understood that any other parameters occurring in the probability law of x are known precisely). Let the required level of significance be α. Then the test, which consists in rejecting H_0 when $x \in R$ and accepting H_0 in all other cases, will be a most powerful test if the critical region R satisfies the two conditions:

$$\left.\begin{aligned} &\text{(i)} \quad \int_{(R)} dF(x \mid \theta_o) = \alpha , \\ &\text{(ii)} \quad \int_{(R)} dF(x \mid \theta_1) = \text{minimum} \end{aligned}\right\} \qquad (1.5)$$

Neyman and Pearson (1933) proved that if a region R exists which satisfies (i) and is such that x belongs to R whenever

$f(x \mid \theta_o) / f(x \mid \theta_1) < c$, where c is some constant,

and such that x does not belong to R whenever $f(x \mid \theta_o) / f(x \mid \theta_1) > c$,

and if also R* is any other region for which $\int_{(R^*)} dF(x \mid \theta_o) \leq \alpha$,

then $\int_{(R^*)} dF(x \mid \theta_1) \leq \int_{(R)} dF(x \mid \theta_1)$.

This means that the test using R is a most powerful one of size α. Unfortunately, a region with the stated properties may not exist. This situation is likely to crop up when the distribution is discrete. An example where the region does exist will be given in § 2.1.

The ratio $f(x \mid \theta_o)/f(x \mid \theta_1)$ is called the likelihood ratio. More generally, if the possible values of θ form a set denoted by Ω and if the null hypothesis H_o implies that θ belongs to a certain sub-set ω of Ω (in symbols, $\theta \in \omega$), the likelihood ratio L(x) is the ratio of the maximum value of f(x) under H_o to the maximum value under H_1, i.e.

$$L(x) = \max_{\theta \in \omega} f(x \mid \theta) \; / \; \max_{\theta \in \Omega - \omega} f(x \mid \theta) \qquad (1.6)$$

If L(x) is small, the observed x is much more likely under H_1 than it is under H_o, so that it would be unreasonable to maintain H_o. The likelihood ratio test consists in rejecting H_o when $L(x) < c$,

c being a constant so chosen that

$$\Pr \left\{ L(x) < c \mid H_o \right\} = \alpha . \tag{1.7}$$

Most of the good tests known are likelihood ratio tests. In many practical cases the statistic X is a set of N independent observations forming a random sample. It has been proved (S. Wilks, 1938) that when N is large the distribution of - 2 log L(x) is approximately an ordinary χ^2 distribution when H_o is true and a non-central χ^2 distribution when H_1 is true. The size and power of the test can be readily calculated from tables. Tables of χ^2 are readily available (e. g. Fisher and Yates, 1949 and C. M. Thompson, 1941-42), and Evelyn Fix (1949) has published a table of non-central χ^2.

1.8 The Randomized Neyman-Pearson Lemma

We can generalize the conditions (1.5) to read

$$\int_{(R)} dF(t \mid \theta) \leq \alpha \tag{1.8}$$

for all $\theta \in \omega$, and

$$\int_{(R)} dF(t \mid \theta) = \max. \tag{1.9}$$

for all $\theta \in \Omega - \omega$ Here $F(t \mid \theta)$ is the cumulative distribution function of the statistic T and θ is the parameter being estimated. The region of rejection R is chosen to satisfy (1.8) and (1.9). According to (1.8) the size of the test is not greater than α .

A possibility of increasing the power is afforded by allowing randomized decisions, as suggested by Lehmann and Stein, 1948. The total space of the statistic T is divided into three parts, R_1, R_2 and R_3. If T falls into R_1, H_o is rejected; if in R_3, H_o is accepted; but if T falls in R_2, we toss a coin or look up in a table of random numbers to decide whether to accept or reject H_o. In other words, we reject H_o with probability $\psi(T)$, whatever the value of T, $\psi(T)$ being 1 for $T \in R_1$, 0 for $T \in R_3$, and a number between 0 and 1 for $T \in R_2$. $\psi(T)$ is called a test function, or sometimes simply a test.

The randomized Neyman-Pearson lemma states that if L(x) is defined as in (1.6) and if a test function $\psi(x)$ is defined by

$$\left.\begin{aligned} \psi(x) &= 1, \text{ when } L(x) < c, \\ \psi(x) &= 0, \text{ when } L(x) > c, \\ \psi(x) &= \psi_0 \text{ when } L(x) = c \ , \end{aligned}\right\} \tag{1.10}$$

then the test which consists in rejecting H_0 with probability $\psi(x)$ is most powerful of size α for testing H_0 against H_1. The value of c is given by

$$\begin{aligned} &\Pr\{L(x) < c \mid H_0\} \leq \alpha \ , \text{ and } \quad \psi_0(x) \text{ by} \\ &\Pr\{L(x) < c \mid H_0\} + \psi_0(x) \Pr\{L(x) = c \mid H_0\} = \alpha . \end{aligned} \tag{1.11}$$

If the region R_2, where $L(x) = c$, contains only a single value of x, $\psi_0(x)$ is unique, but if R_2 includes more than one value of x it will in general be possible to choose different functions satisfying (1.11).

1.9 Examples of Randomized Tests

Suppose we want to test the hypothesis that the proportion of defectives in a certain manufactured article subject to inspection is equal to or less than 10%, and that we want to do so on a sample of 4 items by noting the number x of defective articles in the sample. Clearly, x can take only the values of 0, 1, 2, 3 or 4, and the larger values of x will lead to rejection of the hypothesis.

If the true proportion of defective articles is θ, the probability of exactly x defectives in a sample of 4 is

$$p(x \mid \theta) = \binom{4}{x} \theta^x (1 - \theta)^{4-x} \tag{1.12}$$

For $\theta = 0.10$, this expression takes for $x = 2, 3, 4$, the values 0.0486, 0.0036, 0.0001 respectively, and still smaller values for $\theta < 0.10$. The statistic T is here identical with x itself, so that if we take the region of rejection as R_1 ($x = 3$ or 4), we have

$$\sum_{(R_1)} p(x \mid \quad \theta) \leq 0.0037$$

for all $\theta \leq 0.10$. If, on the other hand, we include also $R_2(x = 2)$,

$$\sum_{(R_1 + R_2)} p(x \mid \theta) \leq 0.0523$$

and so the size of the test is greater than 0.05. The power of the non-randomized test is

$$P(\theta) = \sum_{(R_1)} p(x \mid \quad \theta) = \theta^4 + 4\theta^3 (1 - \theta)$$

for $\theta > 0.10$. For $\theta = 0.20$, this is 0.027 and for $\theta = 0.30$ it is 0.084.

Now let us consider a randomized test, in which, when $x = 2$, we reject H_o with probability ψ_o, where

$$\Pr \{ x > 2 \mid H_o \} + \psi_0 \cdot \Pr \{ x = c \mid H_o \} = .05, \qquad (1.13)$$

i.e. $.0037 + 0.0486\, \psi_o = 0.05$, or $\psi_o = 0.95$.
The power of the test is

$$P(\theta) = \sum \psi(x)\, p(x \mid \quad \theta)$$

where $\psi(x) = 0$ for $x = 0$ or 1, 0.95 for $x = 2$ and 1 for $x = 3$ or 4. Hence $P(\theta) = \theta^4 + 4\theta^3 (1 - \theta) + 0.95\, \theta^2 (1 - \theta)^2$, which is equal to 0.173 for $\theta = 0.20$ and to 0.335 for $\theta = 0.30$.

It was unnecessary in the above work to determine the value of the likelihood ratio $L(x)$. However, it is easily seen, by maximizing $\theta^x (1 - \theta)^{4-x}$, that $L(x) = (10/9)^4$ when $x = 0$ and $L(x) = (0.1)^x (0.9)^4 / \left[(x/4)^x (3x/4)^{4-x} \right]$ when $x > 0$. The value for $x = 2$ is 0.0144, which is the c of (1.11). The probability that $L(x) = c$ is the same as the probability that $x = 2$.

A practical method of rejecting H_o with probability 0.95 would be to use a table of random 2-digit numbers. Before opening the table, decide on a particular page, a particular column, and a particular position in the column (say the 7th from the top). Then look up the corresponding number, and if it lies between 0 0 and 94 inclusive, reject the hypothesis.

In the following example, $\psi_0(x)$ is not unique. Let the null hypothesis H_0 be that x is a random item from a rectangular distribution of mean 2 and range 2, and let the alternative hypothesis H_1 be that x is from a rectangular distribution of mean 4 and range 4 (see Fig. 3).

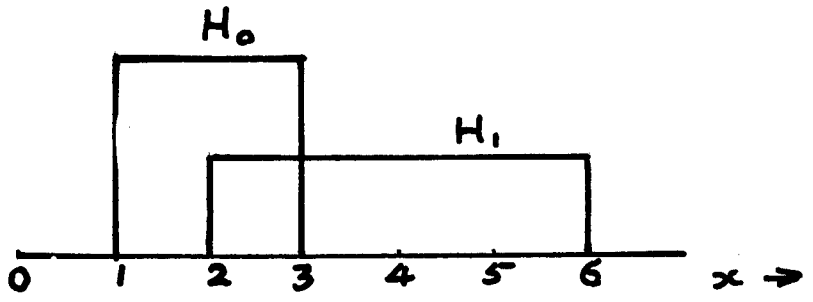


Fig. 3 Two Rectangular Distributions

The null hypothesis must obviously be accepted if $1 < x < 2$ and rejected if $3 < x < 6$. The only doubt arises when x lies between 2 and 3 in the region where the two distributions overlap.

Clearly, $L(x) = \infty, \quad 1 < x < 2,$
$L(x) = 2, \quad 2 < x < 3,$
$L(x) = 0, \quad 3 < x < 6.$

The probability ψ_0 (taken as constant for x between 2 and 3) is here given by

$$\psi_0 = \frac{\alpha - \Pr\{3 < x < 6 \mid H_0\}}{\Pr\{2 < x < 3 \mid H_0\}}$$

$$= \alpha/(1/2) = 0.1$$

if $\alpha = 0.05$

The power of the test is P, where

$$P = \Pr\{3 < x < 6 \mid H_1\} \cdot 1 + \Pr\{2 < x < 3 \mid H_1\} \cdot \psi_0$$

$$= \frac{3}{4} + \frac{1}{4} \cdot \frac{1}{10} = \frac{31}{40} .$$

Another possible value of ψ_0 would be

$$\psi_0(x) = (x - 2)/5, \quad 2 < x < 3,$$

and this may easily be shown to give the same size and power as

$$\psi_0(x) = 0.1.$$

1.10 Similar Tests and Unbiased Tests

A test $\psi(x)$ will be most powerful for a given size α, according to the criteria stated above, if the following conditions are fulfilled:

$$\int \psi(x)\, dF(x|\theta) \leq \alpha \text{ for } \theta \in \omega \tag{1.14}$$

$$\int \psi(x)\, dF(x|\theta) = \text{max. for } \theta \in \Omega - \omega \tag{1.15}$$

That is to say, the probability of rejection of the null hypothesis when this hypothesis is true (i.e. when θ belongs to the set ω) is never greater than α, and the probability of rejection when this hypothesis is false is as great as possible.

The test is said to be similar if strict equality holds in equation (1.14). Neyman and Pearson (1928) originally considered only similar tests, but it is not always convenient to limit oneself to this case, as it puts a considerable restriction on the hypotheses available for testing. Thus, if $X_1, X_2, \ldots, X_n$ are independent normal variates with mean μ and variance σ^2, and we want to test whether $\mu > 0$, we can get a similar test of H_0 $(\mu = 0)$ against H_1 $(\mu > 0)$ but not of H_0 $(\mu \leq 0)$ against H_1 $(\mu > 0)$. The latter case is, however, at least as worth while investigating as the former.

For this reason, another criterion of a good test was introduced by Neyman and Pearson, that of being unbiased. The

test $\psi(X)$ is unbiased if (a) $\int \psi(x)\, dF(x \mid \theta) \leq \alpha$, for $\theta \in \omega$

(b) $\int \psi(x)\, dF(x \mid \theta) \geq \alpha$, for $\theta \in \Omega - \omega$.

That is, the probability of rejecting the null hypothesis is at least as great when the hypothesis is false as when it is true. This is obviously a desirable characteristic of a test. For a simple hypothesis $\theta = \theta_o$, the power curve of an unbiased test of this hypothesis has an absolute minimum at $\theta = \theta_o$. If the curve has a minimum at $\theta = \theta_o$, in the neighborhood of θ_o, but falls below the value corresponding to θ_o at one or more points distant from θ_o, the test is said to be locally unbiased.

1.11 Sufficient Statistics

Given a set of random variables $X_1, X_2, \ldots, X_N$ from a population which has a distribution characterized by a parameter θ (which may be one of a set of parameters), we can use certain functions of the X_i (or of some of them) to estimate θ. Thus, if θ is the mean of the population we can estimate it by the arithmetic mean of all the variables, or by the median, or by half the sum of the smallest and the largest, or in many other ways, but not all these ways are equally good. Some of them, although they may be quick, fail to utilize all the available information. R. A. Fisher defined a sufficient statistic as one containing all the relevant information in the sample. Thus if the X_i are independent and come from a normal population of unknown mean μ and known variance σ^2, it is possible to change the variables to a new independent set $Y_1, Y_2, \ldots Y_N$, which are linearly related to the X's and are such that Y_1 is normal with mean $N^{1/2}\mu$ and variance σ^2, while $Y_2, Y_3, \ldots, Y_N$ are all normal with mean 0 and variance σ^2. If the hypothesis which we wish to test relates to μ, it is clear that $Y_2, Y_3, \ldots, Y_N$ contribute no information, and all we need worry about is Y_1. It is easily verified that the following set of Y's satisfies the required conditions:

$$Y_1 = (X_1 + X_2 + \ldots + X_n)/\sqrt{N}$$
$$Y_2 = (X_1 - X_2)/\sqrt{2}$$
$$Y_3 = (X_1 + X_2 - 2X_3)/\sqrt{6}$$
$$Y_N = (X_1 + X_2 + \ldots + X_{N-1} - (N-1)X_N)/\sqrt{N(N-1)}$$

and Y_1 is simply the arithmetic mean of the X's, multiplied by $N^{1/2}$. The arithmetic mean is therefore a sufficient statistic for estimating μ.

A more precise definition is the following: The statistic T is sufficient for estimating θ if for any other statistic T' the conditional probability (or probability density, in the case of a continuous variable) of T', given T, is independent of θ. The probability of the observed sample is then given by

$$p(X, \theta) = p(T, \theta) \cdot g(X), \tag{1.16}$$

where $p(T, \theta)$ is the probability of the statistic T and g (X) is a function of the sample values which is independent of θ. Hence, if $L = \log p$,

$$L = L_1(T, \theta) + L_2, \tag{1.17}$$

where L_2 is independent of θ, so that a knowledge of L_1 will give all the information about θ obtainable from the sample.

The condition of sufficiency does not determine T completely. Any function of T is also sufficient. We naturally choose a function which gives a consistent estimate of θ, and if possible one that is unbiased. In problems of testing hypotheses we can restrict ourselves to sufficient statistics, because of the theorem that if $\psi(X)$ is any test and if T is a sufficient statistic, it is possible to find a test $\psi'(T)$, depending on T only, which has the same power function as $\psi(X)$ and so is equivalent to $\psi(X)$. This test is in fact defined by

$$\psi'(t) = \int \{\psi(x) \mid t\} \, dF(x \mid \theta) \tag{1.18}$$

where t is any observed value of T.

TESTING THE MEAN OF A SAMPLE FROM A NORMAL POPULATION OF KNOWN VARIANCE.

2.1 Simple Hypothesis against Simple Alternative (One-sided)

Let $X_1, X_2, \ldots, X_N$ be a set of N independent observations of a variable which is normally distributed in the population with unknown mean μ and known variance σ^2. At first sight, this seems rather an unreasonable assumption to make, but it is not without justification in some circumstances. It may happen that conditions affecting this variable have changed in such a way that the mean value

is pushed up or down while the amount of variation about the mean is substantially the same as before. The variance may then be estimated with considerable accuracy from previous observation, but the new mean is not known.

Let the null hypothesis H_0 be that $\mu = \mu_0$ and the alternative hypothesis H_1 be that $\mu = \mu_1$, where we may suppose $\mu_1 > \mu_0$. The test is therefore of a simple hypothesis against a simple alternative. The statistic used to estimate μ is the arithmetic mean $\bar{X}$, the probability density of which is

$$f(\bar{x} \mid \mu_0) = (N/2\pi\sigma^2)^{1/2} \; e^{-N(\bar{x} - \mu_0)^2 / 2\sigma^2} \tag{2.1}$$

The likelihood ratio described in the Neyman-Pearson Lemma, § 1.7 is therefore

$$\frac{f(\bar{x} \mid \mu_0)}{f(\bar{x} \mid \mu_1)} = \exp\left[-\frac{N}{2\sigma^2}(\mu_1 - \mu_0)(2\bar{x} - \mu_1 - \mu_0)\right] \tag{2.2}$$

so that the condition $f(\bar{x} \mid \mu_0) / f(\bar{x} \mid \mu_1) < c$ is equivalent

to $2\bar{x} - (\mu_1 + \mu_0) > c_1$ or $\bar{x} > c_2$, (2.3)

where c_1 and c_2 are other constants depending on c and on the known values of μ_0, μ_1, σ and N. The size of the test is given by

$$\alpha = \int_{c_2}^{\infty} \left(\frac{N}{2\pi\sigma^2}\right)^{1/2} e^{-N(\bar{x} - \mu_0)^2 / 2\sigma^2} \, d\bar{x}, \tag{2.4}$$

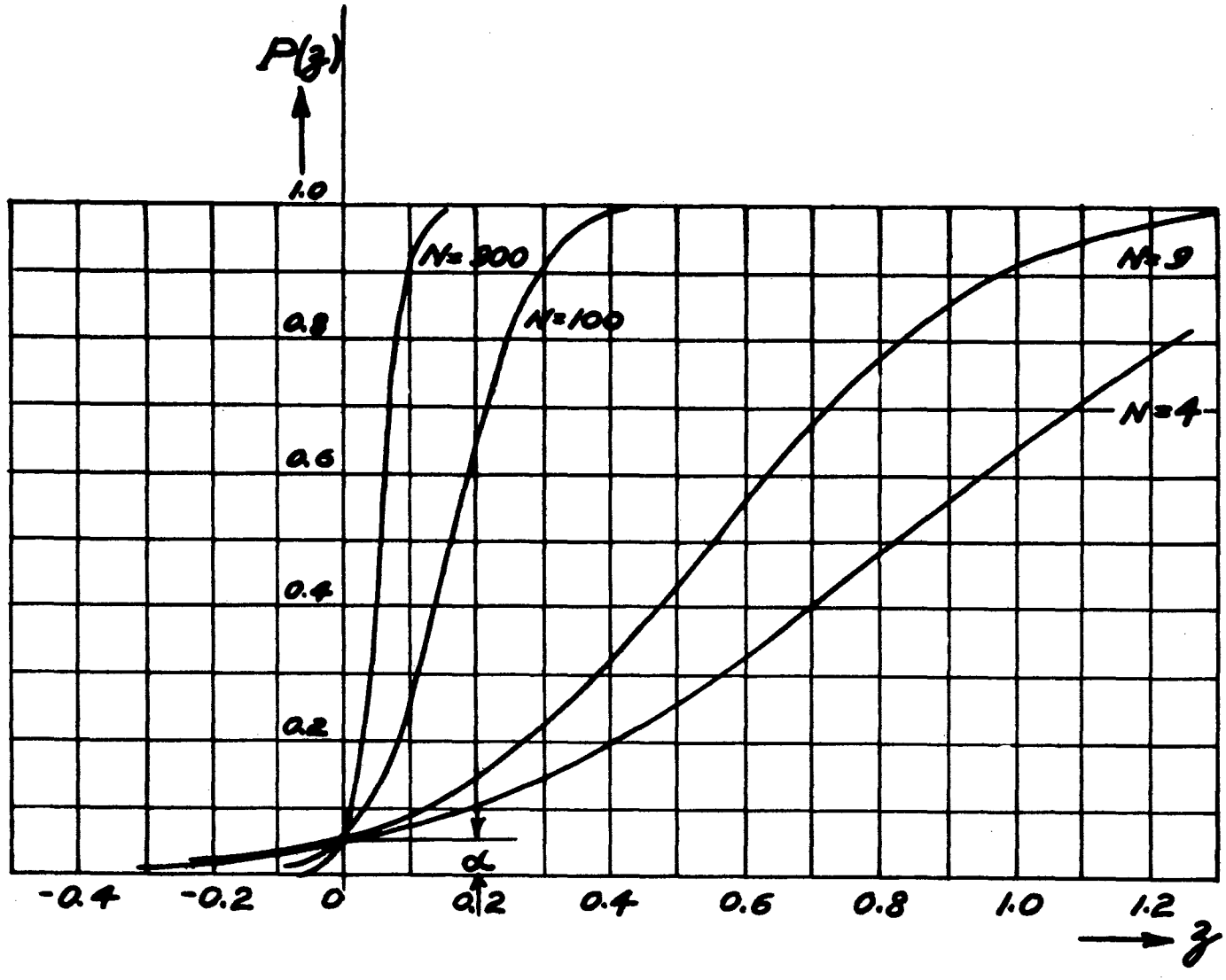
since R_1 is here the interval c_2 to ∞. If $v = N^{1/2}(\bar{x} - \mu_0)/\sigma$, and $v_0 = N^{1/2}(c_2 - \mu_0)/\sigma$, equation (2.4) gives

$$\alpha = 1 - \Phi(v_0), \tag{2.5}$$

where $\Phi(v_0) = \int_{-\infty}^{v_0} \phi(v) \, dv$, and $\phi(v) = (2\pi)^{-1/2} e^{-v^2/2}$,

whence v_0, and therefore c_2 can be obtained. Thus, for $\alpha = 0.05$, $v_0 = 1.645$.

FIG. 4. POWER CURVES for ONE-SIDED NORMAL TEST.



The power of the test is the probability of rejecting H_o when μ is really equal to μ_1, that is, the probability that $\bar{x} > c_2$ when the probability density is given by

$$f(\bar{x}\,|\,\mu_1) = \left(\frac{N}{2\pi\sigma^2}\right)^{1/2} e^{-N(\bar{x}-\mu_1)^2/(2\sigma^2)}$$

The power is therefore

$$P(\chi) = \int_{c_2}^{\infty} f(\bar{x}\,|\,\mu_1)\,d\bar{x} = 1 - \Phi(v_1), \qquad (2.6)$$

where $v_1 = N^{1/2}(c_2 - \mu_1)/\sigma = v_o - (\mu_1-\mu_o)N^{1/2}/\sigma = 1.645 - N^{1/2}z,$

with $z = (\mu_1-\mu_o)/\sigma$.

Some power curves are shown in Fig. 4. For example, if z = 0.3 and N = 9, $P(\chi) = 1 - \Phi(0.745) = 0.228$.

2.2. Size of Sample necessary for Detecting a given Difference in the Mean

The simple example in § 2.1 illustrates how power functions may be used in designing experiments. If the true mean μ_1 differs from the assumed mean μ_o, we may or may not be able to detect this difference by using a sample of size N. Let us suppose that we want to have at least a 100P% chance of detecting a difference equal to z times the standard deviation. Since $P(\chi)$ in equation (2.6) is the probability of rejecting the hypothesis that the mean is μ_o when it is really μ_1, we have

$$P = 1 - \Phi(1.645 - z\,N^{1/2}) \qquad (2.7)$$

Some values of N calculated from this equation, for given values of P and z, are collected in Table I. It appears, for instance, that to have a 90% chance of detecting a difference in the means equal to 0.3 of the standard deviation (with a test which has only a 5% chance of claiming that such a difference exists when in fact there is no difference at all) we need a sample size of at least 96. This assumes that any difference that does exist can only be an increase, but the same result holds if we know that the difference must be a decrease. The two-sided alternative will be discussed in the next section. In the first two lines of Table I, the sample sizes, being large, have been rounded off.

TABLE I

Size of Sample necessary to detect with Probability P a One-sided Difference in the Mean equal to $z\sigma$ (Normal Population).

z \ P	0.9	0.8	0.7	0.6	0.5
0.01	85,700	61,900	47,100	36,100	27,100
0.02	21,410	15,460	11,770	9,010	6,770
0.05	3,426	2,474	1,883	1,442	1,083
0.1	857	619	471	361	271
0.2	215	155	118	91	68
0.3	96	69	53	41	31
0.4	94	39	30	23	17
0.5	35	25	19	15	11
0.6	24	18	14	11	8
0.7	18	13	10	8	6
0.8	14	10	8	6	5
0.9	11	8	6	5	4
1.0	9	7	5	4	3
1.5	4	3	3	2	2

The probability of apparently detecting a difference that does not exist is assumed to 0.05.

2.3 Simple Hypothesis against Two-sided Alternative

Let the null hypothesis H_o be that $\mu = \mu_o$, and the alternative hypothesis H_1 that $\mu > \mu_o$ or $\mu < \mu_o$. Let us also agree to reject H_o if either $\bar{x} - \mu_o > c$ or $\mu_o - \bar{x} > c$, where c is a fixed value, determined by the size α of the test. By (1.1), we have

$$\int_{-\infty}^{\mu_o - c} f(\bar{x} \mid \mu_o) d\bar{x} + \int_{\mu_o + c}^{\infty} f(\bar{x} \mid \mu_o) \, d\bar{x} = \alpha \quad , \tag{2.8}$$

where $f(\bar{x} \mid \mu_o)$ is given by (2.1).

Due to the symmetry of the distribution, both integrals are equal to $\alpha/2$.

Putting $v = N^{1/2} (\bar{x} - \mu_o)/\sigma$, $\phi(v) = (2\pi)^{-1/2} \, e^{-v^2/2}$,

and $\bar{\Phi}(u) = \int_{-\infty}^{u} \phi(v) dv$, we find

$$\int_{c N^{1/2}/\sigma}^{\infty} \phi(v) dv = 1 - \bar{\Phi}(c N^{1/2} / \sigma) = \alpha/2, \tag{2.9}$$

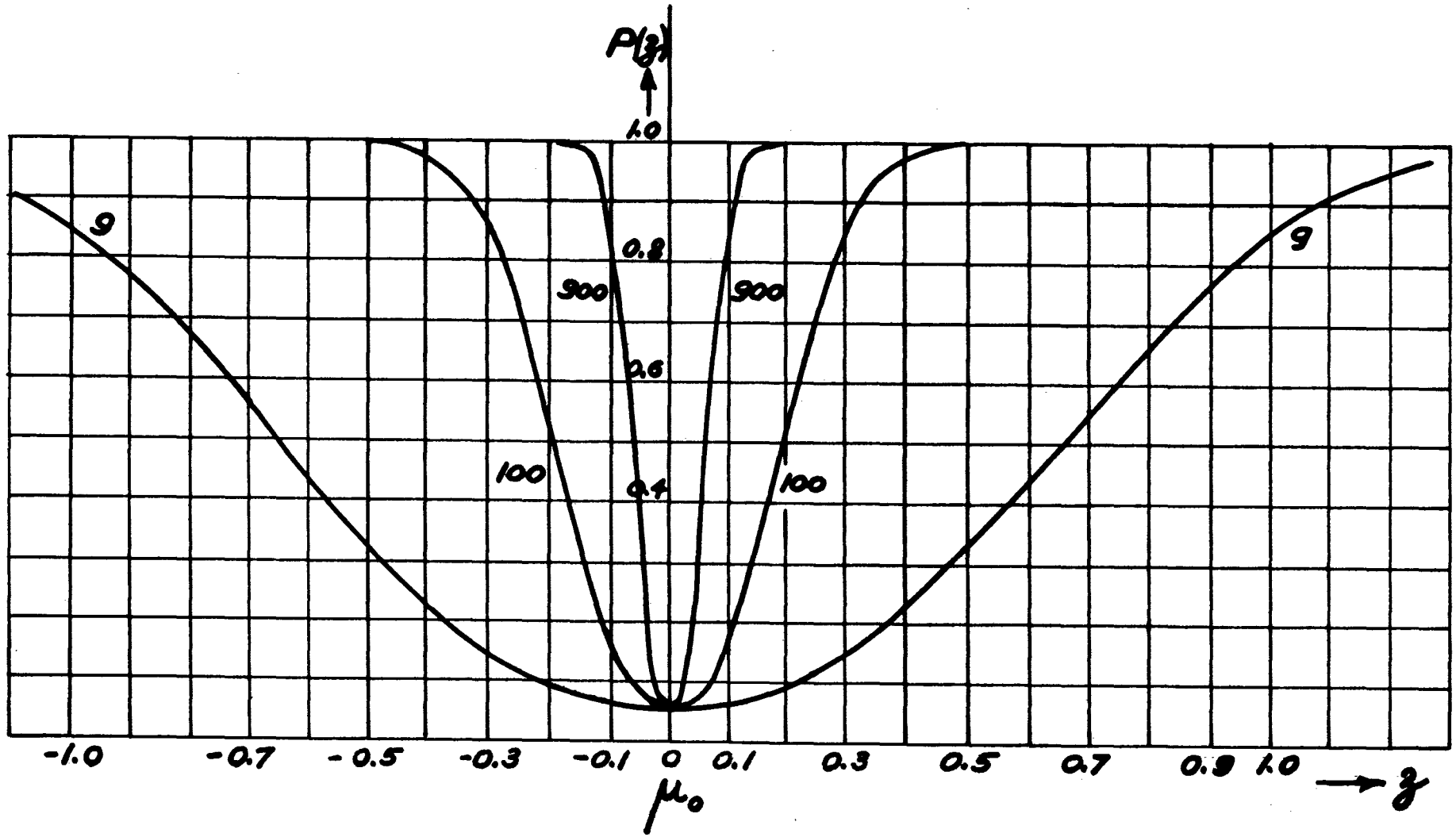
so that c is known when α has been fixed. If, for example, $\alpha = 0.05$, we readily find from tables of the normal law that $\bar{\Phi}(c N^{1/2} / \sigma) = 0.975$, whence $c = 1.96 \, \sigma N^{-1/2}$.

The power of the test, by (1.2), is given by

$$P(\mu) = 1 - \int_{\mu_o - c}^{\mu_o + c} f(\bar{x} \mid \mu) \, d\bar{x}$$

$$= 1 - \bar{\Phi}\left[1.96 - N^{1/2} \sigma^{-1} (\mu - \mu_o)\right] + \bar{\Phi}\left[-1.96 - N^{1/2} \sigma^{-1} (\mu - \mu_o)\right]$$

FIG. 5 POWER CURVES for TWO-SIDED NORMAL TEST.



If $z = (\mu - \mu_0)/\sigma$, the power function in terms of z is

$$P(z) = 1 - \Phi\,(1.96 - N^{1/2} z) + \Phi\,(-1.96 - N^{1/2} z) \qquad (2.10)$$

This function is drawn, for a few selected values of N, in Figure 5. The curves show clearly how, as the sample size increases, the test approximates more and more closely to the ideal test pictured in Fig. 1.

Figure 5 also shows that for a sample of 9 the test is not very powerful in rejecting a false hypothesis. Even when the true population mean and the hypothetical mean differ by a whole standard deviation, the chance of not detecting this discrepancy is about 0.15.

We now prove that this test is identical with the likelihood ratio test.

The probability of the observed set of N sample values, $x_1, x_2, \ldots, x_N$, on the assumption that the true mean is μ is

$$(2\pi\sigma^2)^{-N/2} \exp\left[-\sum_i (x_i - \mu)^2 / 2\sigma^2 \right]$$

and this is the probability that is to be maximized in (1.6).

Since in this example ω consists of the single value μ_0, the numerator is simply

$$(2\pi\sigma^2)^{-N/2} \exp\left[-\sum_i (x_i - \mu_0)^2 / 2\sigma^2 \right]$$

The denominator is a maximum when $\mu = \bar{x}$, so that

$$L = \frac{\exp\left[-\sum (x_i - \mu_0)^2 / 2\sigma^2 \right]}{\exp\left[-\sum (x_i - \bar{x})^2 / 2\sigma^2 \right]}$$

or $\log L = \left[\sum (x_i - \bar{x})^2 - \sum (x_i - \mu_o)^2 \right] / 2\sigma^2$. (2.11)

The condition $-2 \log L > c_1^2$ corresponds therefore to $(\bar{x} - \mu_o)^2 > c_1^2 \sigma^2/N$, so that $\bar{x} - \mu_o > c_1 \sigma N^{-1/2}$ or $\bar{x} - \mu_o < - c_1 \sigma N^{-1/2}$ and so is the same as the test given above with $c_1 \sigma N^{-1/2} = c$. To say that $-2 \log L > c_1^2$ is of course the same as to say that $L < e^{-c_1^2/2}$.

Since $-2 \log L = N(\bar{x} - \mu_o)^2 / \sigma^2$, which on the null hypothesis is the square of a standard normal variate, it follows that its distribution is that of χ^2 with 1 degree of freedom. The asymptotic approximation mentioned in § 1.7 is in this case exact for all N.

2.4 Size of Sample necessary for Detecting a given Difference in the Mean with Two-sided Alternative.

Since P(z) in equation (2.10) is the probability of rejecting the hypothesis that the mean is μ_o when it is really μ, we must take P(z) as given, say 0.80. Then

$$\Phi(1.96 - z N^{1/2}) - \Phi(-1.96 - z N^{1/2}) = 0.2 \quad . \tag{2.12}$$

The same numerical values occur whether $z > 0$ or < 0, because of the symmetry of the normal curve. It is clear from Fig 5 that for $z = 0.1$ the value of N will be near to 900 (the ordinate at $z = 0.1$ and the abscissa through $P(z) = 0.8$ meet at a point between the curves for $N = 100$ and $N = 900$, but much nearer to the latter). Similarly, for $z = 0.3$, N is a little under 100, and for $z = 1$, it is a little under 9. Hence $z N^{1/2}$ is approximately equal to 3, and $\Phi(-1.96 - z N^{1/2})$ is therefore practically zero. Since $\Phi(-0.8416) = 0.2$, equation (2.12) is approximately equivalent to $1.96 - z N^{1/2} = -0.8416$, or $N = (2.802/z)^2$, whence N is readily calculated for any given z. Similarly for a power 0.90, we should have $N = (3.242/z)^2$ and for a power 0.50, $N = (1.960/z)^2$. These approximations may be checked by substituting in the exact equation (2.12). We thus arrive at the values given in Table II, for sample sizes necessary to detect a difference of $z\sigma$ in the mean with the

TABLE II

Size of Sample necessary to detect with Probability P a Difference either way of $\chi\,\sigma$ in the Mean. (Normal Population).

χ \ P	0.9	0.8	0.7	0.6	0.5
0.01	105,200	78,600	61,800	49,000	38,500
0.02	26,280	19,630	15,430	12,250	9,610
0.05	4,204	3,140	2,469	1,960	1,537
0.1	1,051	785	618	490	385
0.2	263	197	155	123	97
0.3	117	88	69	55	43
0.4	75	50	39	31	25
0.5	43	32	25	20	16
0.6	30	22	18	14	11
0.7	22	17	13	10	8
0.8	17	13	10	8	7
0.9	13	10	8	7	5
1.0	11	8	7	5	4
1.5	5	4	3	3	2

The probability of stating that a difference exists, when in fact, it does not, is supposed to 0.05 throughout.

assigned probabilities, the probability of falsely rejecting the null hypothesis being assumed to be 0.05. Since N must be integral, the left-hand side of equation (2.12) will actually be ≤ 0.2, but N is the smallest value for which this inequality holds.

For the same values of z and P, the value of N given by Table II is greater than that given in Table I. This is because we no longer know that the difference to be looked for is in a given direction.

2.5 Simple Hypothesis against Composite Alternative

The problem here is to find the distribution under H_1, which may be different for different values of the parameter θ. Suppose, for instance, that H_0 is the hypothesis that a sample of N observations has been drawn from a normal population of mean μ_0 and variance σ^2. The alternative hypothesis is that the mean is μ and variance σ^2, where $\mu \neq \mu_0$.

On the null hypothesis the quantity $\sum_i (x_i - \mu_0)^2 / \sigma^2$ is distributed as χ^2 with N degrees of freedom. If $f_N(\chi^2)$ is the probability density for χ^2, and if $\chi^2_{N,\alpha}$ is defined by

$$\Pr \left\{ \chi^2 > \chi^2_{N,\alpha} \right\} = \alpha , \tag{2.13}$$

an unbiased test of H_0, of size α, is given by the rule of rejecting H_0 when

$$\sum (x_i - \mu_0)^2 > \sigma^2 \chi^2_{N,\alpha} . \tag{2.14}$$

Under the alternative hypothesis, the quantity $x = \sum (x_i - \mu_0)^2 / \sigma^2$ follows the non-central χ^2 distribution, with N degrees of freedom. This distribution depends on a parameter

$$\lambda = N(\mu - \mu_0)^2 / \sigma^2 \tag{2.15}$$

and the probability density of x is

$$f(x) = 1/2\, e^{-\lambda/2} (x/2)^{\frac{N-2}{2}} e^{-x/2} \sum_{m=0}^{\infty} \frac{(\lambda x/4)^m}{m!\, \Gamma(m+1/2\, N)} \tag{2.16}$$

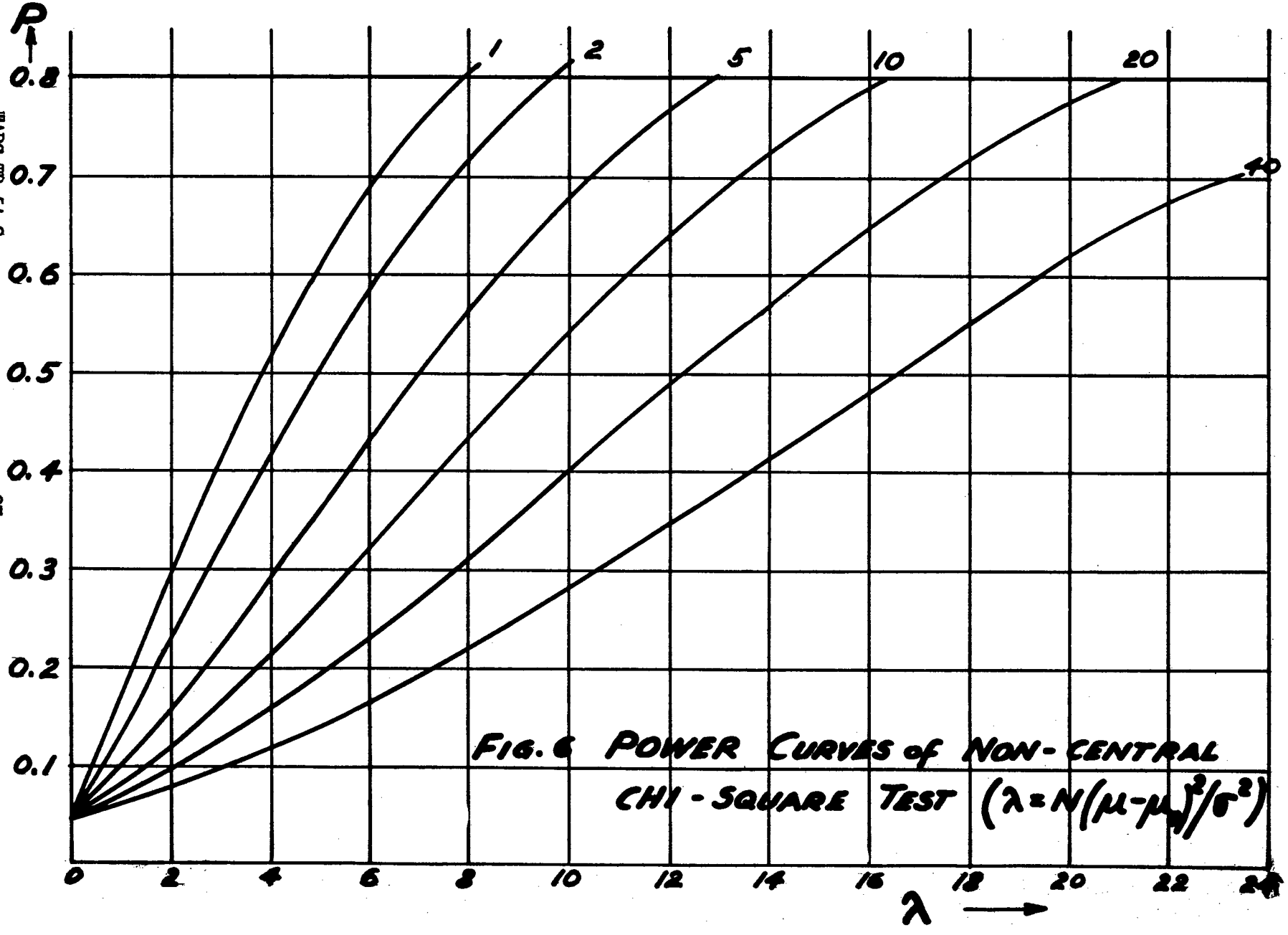


FIG. 6 POWER CURVES of NON-CENTRAL CHI-SQUARE TEST $(\lambda = N(\mu-\mu_0)^2/\sigma^2)$

When $\lambda = 0$, f(x) reduces to the ordinary χ^2 distribution.

The power of the test is

$$P = \Pr\left\{ x > \chi^2_{N,\alpha} \middle| H_1 \right\} \qquad (2.17)$$

$$= \int_{\chi^2_{N,\alpha}}^{\infty} f(x)dx$$

Numerical tables of non-central χ^2, giving λ for specified values of P, have been prepared by E. Fix (1949). Curves derived from these tables for $\alpha = 0.05$ are given in Fig. 6. If, for example, $N = 10$, $\mu_0 = 0$, $\mu = 0.5$, and $\sigma = 1$, we have $\lambda = 2.25$. It is evident from the curve that the power of the test is about 0.13. If N were 40, λ would be 10 and the power would be 0.28.

2.6 Composite Hypothesis against Simple Alternative

The general problem of testing a composite null hypothesis H_0 ($\theta \in \omega$) against a simple alternative H_1 ($\theta = \theta_1$ where $\theta_1 \in \Omega - \omega$) has been considered by Lehmann and Stein, 1948.

Let us denote the distribution function of x under H_1 by G(x) and that under H_0 by $F_\theta(x)$, to indicate that it depends on the parameter θ. The procedure adopted by Lehmann and Stein consists in replacing the composite hypothesis H_0 by a simple hypothesis H_0', this hypothesis being that θ has a particular distribution over ω, so chosen that the problem of distinguishing between H_0' and H_1 is as difficult as possible. If we then find a most powerful test for H_0' against H_1 it turns out to be, under certain conditions, also most powerful for H_0 against H_1.

It is often possible to guess the least favorable distribution. With the same assumption as in § 2.1, let the null hypothesis be that $\mu \leq \mu_0$ and the alternative hypothesis that $\mu = \mu_1$, where $\mu_1 > \mu_0$. It seems fairly obvious that it will be more difficult to distinguish between μ and μ_1 when $\mu = \mu_0$ than when $\mu < \mu_0$. The least favourable distribution of μ will therefore be a concentration at $\mu = \mu_0$. In other words, the distribution function of μ is a step function with a single

step of height 1 at μ_0, so that the simple null hypothesis H_o' is that $\mu = \mu_o$. The problem is reduced to that of § 2.1.

The most powerful size α test of H_o' against H_1 is to reject H_o' with probability $\psi(\bar{x})$, where

$$\left.\begin{aligned} \psi(\bar{x}) &= 1, \text{ when } \bar{x} > c, \\ \psi(\bar{x}) &= 0, \text{ when } \bar{x} < c, \end{aligned}\right\} \qquad (2.18)$$

c being determined by

$$\Pr\{\bar{x} > c \mid \mu_o\} = \alpha .$$

Now if $\Pr\{\bar{x} > c \mid \mu\} \leq \alpha$ for all $\mu \leq \mu_o$, the test for H_o' against H_1 is also most powerful for H_o against H_1. But this is true, because, as is evident from Fig. 7,

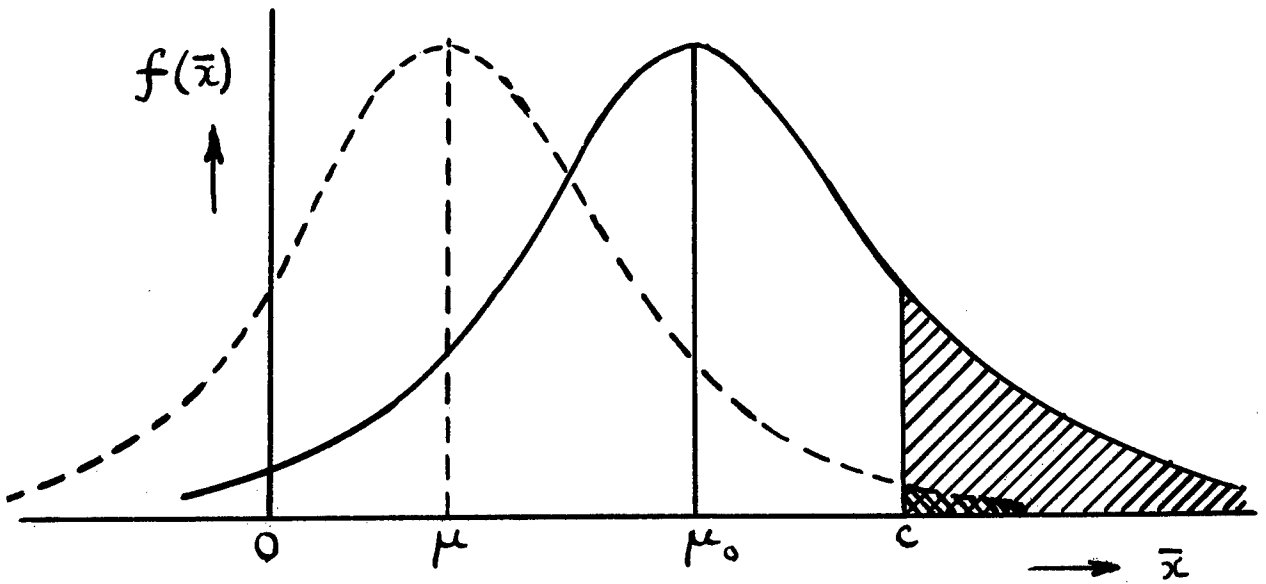


Fig. 7

the area beyond $\bar{x} = c$ gets smaller and smaller as μ moves further to the left from μ_o. The test function (2.18) is therefore most powerful for H_o against H_1, and since it does not depend on the value of μ_1,

it is a uniformly most powerful test of the hypothesis $\mu \leq \mu_0$ against the composite alternative hypothesis $\mu > \mu_0$.

Again, let us suppose that we want to test the composite hypothesis that either $\mu \geq \mu_0$ or $\mu \leq -\mu_0$ against the alternative simple hypothesis that $\mu = 0$. It seems intuitively obvious that the least favorable distribution of μ for the test would be given by a probability of 1/2 for each of the values μ_0 and $-\mu_0$. On this assumption we have

$$f(\bar{x}|H_0') = f'(\bar{x}) = \frac{1}{2}\left(\frac{N}{2\pi\sigma^2}\right)^{1/2}\left[e^{-\frac{N(\bar{x}-\mu_0)^2}{2\sigma^2}} + e^{-\frac{N(\bar{x}+\mu_0)^2}{2\sigma^2}}\right]$$

$$f(\bar{x}|H_1) = g(\bar{x}) = \left(\frac{N}{2\pi\sigma^2}\right)^{1/2} e^{-\frac{N\bar{x}^2}{2\sigma^2}} \qquad (2.19)$$

The most powerful test of size α for H_0' is given by putting

$$\left.\begin{aligned} \psi(\bar{x}) &= 1 \text{ when } f'(\bar{x}) < cg(\bar{x}), \\ \psi(\bar{x}) &= 0 \text{ when } f'(\bar{x}) > cg(\bar{x}), \end{aligned}\right\} \qquad (2.20)$$

where c is determined by the size of the test, namely

$$\Pr\left\{f'(\bar{x}) < cg(\bar{x}) \,\middle|\, H_0'\right\} = \alpha. \qquad (2.21)$$

Now the first condition of (2.20) is equivalent to: $\psi(\bar{x}) = 1$ when

$$\frac{1}{2} e^{N\bar{x}^2/2\sigma^2}\left[e^{-\frac{N(\bar{x}-\mu_0)^2}{2\sigma^2}} + e^{-\frac{N(\bar{x}+\mu_0)^2}{2\sigma^2}}\right] < c,$$

that is, when

$$\frac{1}{2} e^{-N\mu_0^2/2\sigma^2}\left[e^{\frac{N\bar{x}\mu_0}{\sigma^2}} + e^{-\frac{N\bar{x}\mu_0}{\sigma^2}}\right] < c,$$

or when

$$\cosh(N\mu_0\bar{x}/\sigma^2) < c_1,$$

where c_1 is another constant.

The condition may therefore be written $|\bar{x}| < c_2$,

where c_2 depends on c_1. The absolute symbol occurs because of the symmetry of the cosh function. $\bar{x}$ may be anywhere between $-c_2$ and $+c_2$. Since c_2 is merely another constant, we can drop the subscript and call it c. This constant c is then determined by

$$\Pr\left\{ |\bar{x}| < c \mid H_o' \right\} = \alpha .$$

But since the distribution function for μ is a step function with steps 1/2 at $-\mu_o$ and μ_o, we have

$$1/2 \Pr\{|\bar{x}| < c \mid \mu = \mu_o\} + 1/2 \Pr\{|\bar{x}| < c \mid \mu = -\mu_o\} = \alpha$$

and this condition will be satisfied by choosing c so that

$$\Pr\{|\bar{x}| < c \mid \mu = \mu_o\} = \Pr\{|\bar{x}| < c \mid \mu = -\mu_o\} = \alpha . \qquad (2.22)$$

The distribution of $\bar{x}$ is normal with mean μ and variance σ^2/N, so that the value of c is determined by

$$\Phi\{N^{1/2}(c - \mu_o)/\sigma\} - \Phi\{N^{1/2}(-c - \mu_o)/\sigma\} = \alpha . \qquad (2.23)$$

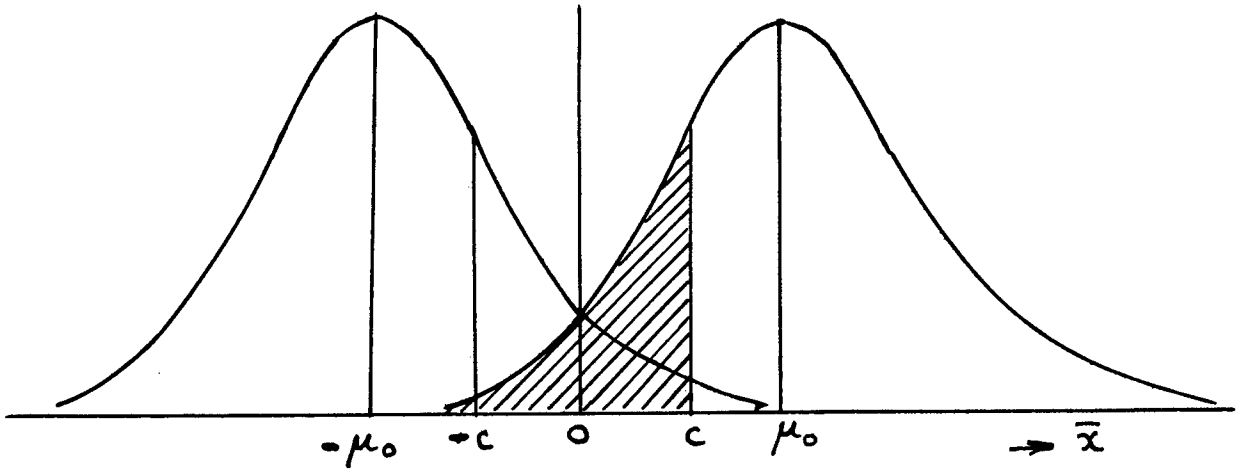


Fig. 8

The shaded area in Fig. 8 (or the equal and symmetrically situated area under the other normal curve) is equal to α. It is easy to see that if $\mu > \mu_0$ or if $\mu < -\mu_0$ the area, for a fixed c, will be less than that shown in the figure, so that

$$\Pr\{\ |\bar{x}| < c|\mu\} = \Pr\{|\bar{x}| < c\,|-\mu\} \leq \alpha$$

for $|\mu| \geq |\mu_0|$. The test given is therefore the most powerful test of size α for testing the hypothesis that $|\mu| \geq |\mu_0|$ against the hypothesis that $\mu = 0$.
The power of the test is given by

$$P = \Pr\{\ |\bar{x}| < c|\mu = 0\}$$
$$= 2\int_0^{cN^{1/2}/\sigma} \phi(v)\, dv\ , \qquad v = \bar{x}\, N^{1/2}/\sigma\ . \tag{2.24}$$

The size of sample necessary to avoid (with probability P) claiming that a difference of the mean from zero at least as great as $z\sigma$ exists, when in fact it does not, is given by solving (2.24) for $cN^{1/2}/\sigma$ and substituing in (2.23) for a given value of α, with $z = \mu_0/\sigma$. Thus for $\alpha = 0.05$, $z = 0.1$, and P = 0.5, we find that N = 532. This greater than the sample size (385) given in Table II for the same values of α, z and P. The reason is that we used Table II to test the null hypothesis $\mu = 0$ against the alternative hypothesis $|\mu| = z\sigma$, whereas now we are testing the null hypothesis $|\mu| = z\sigma$ against the alternative hypothesis $\mu = 0$. In the first case the probability of wrongly claiming that a difference of the mean from zero exists is equal to α, and the power is the probability of detecting a real difference. In the second case the probability of not finding a real difference is equal to α, and the power is the probability of stating that no difference exists when this is in fact true. If we want to be reasonably sure that we do not claim a non-existent effect as real, but do not so much mind missing a real effect, the first arrangement is the one to use and the sample size is given by Table II. If we want to be reasonably sure of finding an effect if it exists, but do not so much mind claiming a non-existent one as real, the second arrangement is better. Some sample sizes for this case are given in Table III.

As an example, suppose we know from previous experience that the width of a slot in a certain metal part is apt to vary, with a standard deviation of 2 thousandths of an inch. If we want to have

an even chance of detecting a real difference of 1 thousandth of an inch in the mean, from an assumed standard value, with a reasonable certainty (95%) that we shall not claim that this difference exists when it really does not, we shall need, according to Table II, a sample of 16. If, on the other hand, we want to be 95% sure of finding this difference if it exists, but are content with a 50% chance of stating that it exists when it really does not, we need a sample of 22 according to Table III.

TABLE III

Size of Sample necessary for Probability P of not finding a Difference of $\chi\sigma$ in the Mean where none actually exists. (Normal Population).

χ \ P	0.9	0.8	0.7	0.6	0.5
0.05	4,329	3,425	2,874	2,465	2,127
0.1	1,083	857	719	617	532
0.2	271	215	180	155	133
0.3	121	96	80	69	60
0.4	68	54	45	39	34
0.5	44	35	29	25	22
0.6	31	24	20	18	15
0.7	21	18	15	13	11
0.8	17	14	12	10	9
0.9	14	11	9	8	7
1.0	11	9	8	7	6

The probability of stating that a difference does not exist, when in fact it does, is supposed to be 0.05 throughout.

111 TESTING THE VARIANCE OF A SAMPLE FROM A NORMAL POPULATION.

3.1 Simple Hypothesis against Simple Alternative

Suppose that the variables X_i (i = 1, 2, ... , N) are independent and normally distributed with mean 0 and variance σ^2, the value of σ^2 being unknown. Suppose also that we wish to test the simple hypothesis H_o (that $\sigma^2 = \sigma_o^2$) against the simple alternative hypothesis H_1 (that $\sigma^2 = \sigma_1^2$).

The statistic used to estimate σ^2 will naturally be the sample variance* v (= s^2) or, to remove bias, $N v/(N - 1)$. The probability density of v, under H_o, is

$$f(v) = \left(\frac{N v}{2\sigma_o^2}\right)^{n/2} \left[v\, \Gamma\left(\frac{n}{2}\right) \right]^{-1} e^{-nv/2\sigma_o^2}$$

where n = N - 1. The probability density g (v) under H_1 is the same expression with σ_1^2 substituted for σ_o^2. Therefore

$$f(v) / g(v) = \left(\sigma_1^2 / \sigma_o^2 \right)^{n/2} e^{\frac{-nv}{2}\left(\frac{1}{\sigma_o^2} - \frac{1}{\sigma_1^2}\right)} \qquad (3.1)$$

First, we suppose that $\sigma_1^2 > \sigma_o^2$. Then the condition $f(v) / g(v) < c$ is equivalent to $v > c$, where the second c is another constant. (In future we shall often use the symbol c for an undetermined constant which may vary from one line to another. This saves writing subscripts such as c_1, c_2, etc.)

The test consists, therefore, in rejecting the hypothesis H_o when $v > c$, c being determined by the condition

$$\Pr\left\{ v > c \,\middle|\, H_o \right\} = \alpha \qquad (3.2)$$

---

* $v = \sum (X_i - \bar{X})^2 / N$. Some writers use N-1 instead of N in defining v, so that v is then an unbiased estimate of σ^2.

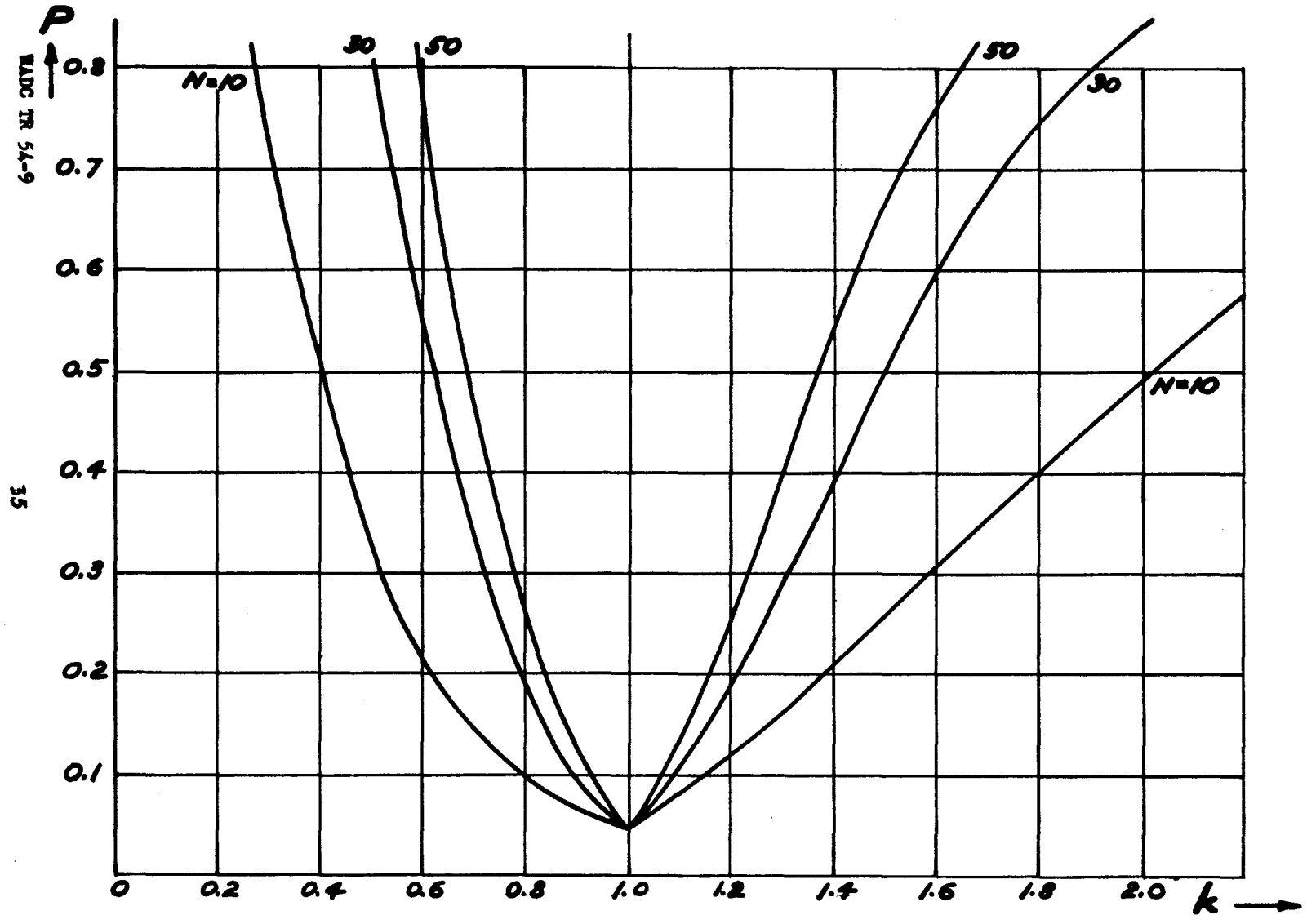


FIG 9. POWER CURVES for the CHI-SQUARE TEST of VARIANCE

$(k = \sigma_1^2 / \sigma_0^2)$

Now when H_o is true, Nv / σ_o^2 has the χ^2 distribution with n degrees of freedom, so that (3.2) is equivalent to

$$\Pr\left\{ \chi_n^2 > Nc / \sigma_o^2 \right\} = \alpha . \tag{3.3}$$

When $\sigma_1^2 < \sigma_o^2$, the exponent in (3.1) is positive so that $f(v) / g(v)$ increases as v increases. The condition $f(v) / g(v) < c$ is equivalent to $v < C$.

The power of the test (for $\sigma_1^2 > \sigma_o^2$) is given by

$$P = \Pr\left\{ v > c \mid H_1 \right\} = \Pr\left\{ \chi_n^2 > Nc / \sigma_1^2 \right\} . \tag{3.4}$$

If $\chi^2_{n,\alpha}$ is such that $\Pr\left\{ \chi_n^2 > \chi^2_{n,\alpha} \right\} = \alpha$,

we have from (3.3) and (3.4) that

$$\chi^2_{n,\alpha} = Nc / \sigma_o^2 , \quad \chi^2_{n,P} = Nc / \sigma_1^2 , \tag{3.3}$$

so that if $k = \sigma_1^2 / \sigma_o^2$,

$$k = \chi^2_{n,\alpha} / \chi^2_{n,P} . \tag{3.5}$$

For given values of k and α, this equation may be solved for P, given n, or for n, given P. The former solution is a straightforward matter of interpolating in the tables of χ^2 (such as those of C. M. Thompson (1941) or A. Hald (1952)). Thus, for $\alpha = 0.05$, and N = 10, $\chi^2_{n,\alpha} = 16.9$. Then, for k = 2, $\chi^2_{n,P} = 8.45$, giving P = 0.49.

Figure 9 gives power curves for a few values of N. When $k < 1$, the inequality signs are reversed, and

$$k = \chi^2_{n,\,1-\alpha} / \chi^2_{n,\,1-P} . \tag{3.6}$$

Thus, for $\alpha = 0.05$ and N = 10, $\chi^2_{n,\,1-\alpha} = 3.325$, and when $k = 0.5$, $P = 0.33$.

In order to find the size of sample necessary to detect a variation in the ratio $\sigma_1^2 / \sigma_o^2 = k$ from the value 1, or in other words, to have a known probability P of rejecting the null hypothesis

$\sigma^2 = \sigma_o^2$ when in fact the true value of σ^2 is $k\ \sigma_o^2$, it is necessary to solve (3.5) or (3.6) for n. This can be most readily done by means of the Wilson and Hilferty (1931) approximation for χ^2, according to which $(\chi^2/n)^{1/3}$ is approximately normal, with mean $1 - 2/(9n)$ and variance $2/(9n)$. Even for n as small as 4, the error in χ^2_α for $\alpha = 0.05$ is less than 1 part in 300. Writing s for $[2/(9n)]^{1/2}$, we see that $(\chi^2_{n,\alpha}/n)^{1/3} - 1 + s^2 = s\, z_\alpha$,

where $\int_{z_\alpha}^{\infty} \phi(z)\, dz = \alpha$, $\phi(z) = (2\pi)^{1/2}\ e^{-z^2/2}$,

and similarly

$$(\chi^2_{n,\,P}/n)^{1/3} - 1 + s^2 = s\, z_p.$$

Therefore, by (3.3)

$$k^{1/3} = (1 - s^2 + s z_\alpha)\ /\ (1 - s^2 + s z_p),$$

and equation which is to be solved for s. If $k < 1$, replace z and z_p by $z_{1-\alpha}$ and z_{1-p} respectively.

A short table of values of N for different values of k and P is given in Table IV. Fuller tables may be found in C. Eisenhart, M. W. Hastay and W. A. Wallis (1947), Chapter 8.

3.2 Different Choice of Test Statistic

Instead of using the sample variance v as the test statistic we may use the whole set of sample values. The probability of the observed sample under H_o is

$$p_o(x) = (2\pi\ \sigma_o^2)^{-N/2}\ e^{-\sum x_i^2/2\,\sigma_o^2}$$

where x stands for the set $x_1, x_2, \ldots, x_N$.

Under H_1 it is

$$p_1(x) = (2\pi\ \sigma_1^2)^{-N/2}\ e^{-\sum x_i^2 / 2\sigma_1^2}.$$

If we agree to reject H_o with probability $\psi(x)$, where $\psi(x) = 1$ when $p_o(x) / p_1(x) < c$, and $\psi(x) = 0$ when $p_o(x) / p_1(x) > c$, the test is equivalent to

$$\psi(x) = 1 \text{ if } \sum x_i^2 > c \tag{3.7}$$

when $\sigma_1^2 > \sigma_o^2$ and

$$\psi(x) = 1 \text{ if } \sum x_i^2 < c \tag{3.8}$$

when $\sigma_1^2 < \sigma_o^2$.

Now $\sum x_i^2 / \sigma_o^2$ is distributed under H_o as χ^2 with N degrees of freedom, not N - 1, so that equations (3.5) and (3.6) still hold with N instead of n. This means that for a given N the power is slightly greater than before, or, to put it another way, the size of sample for a given power, as determined from Table IV, may be reduced by 1. The reason for this is that we are utilizing the known fact that the mean of the population is zero, whereas the test using the variance does not use this information.

3.3 Composite Hypothesis against Simple Alternative

Suppose that the random variables X_i are independently and normally distributed with mean μ and variance σ^2, neither of which is known. Let the null hypothesis H_o be that $\sigma = \sigma_o$, nothing being said about the value of μ, and let the alternative hypothesis H_1 be that $\sigma = \sigma_1$ and $\mu = \mu_1$.

If $\sigma_1 < \sigma_o$, it seems reasonable to assume that the difficulty of distinguishing between H_o and H_1 would be as great as possible if we took $\mu = \mu_1$. Any probability distribution of μ over values other than μ_1 could only increase the variance σ_o and make it still

TABLE IV

Size of Sample necessary to detect with Probability P a Variance Ratio k different from 1.

$k = \sigma_1^2/\sigma_0^2$ \ P	0.5	0.6	0.7	0.8	0.9
0.2	5	6	7	8	9
0.4	10	13	15	18	23
0.5	16	20	24	30	39
0.6	26	33	41	52	69
0.7	50	64	80	102	139
0.8	119	155	198	257	349
0.9	507	668	866	1,130	1,552
1.1	580	778	1,016	1,349	1,878
1.2	154	209	275	366	513
1.3	74	99	134	176	248
1.5	31	42	55	74	105
1.7	18	24	33	44	63
1.9	12	17	22	30	43
2.0	11	15	19	26	37
2.5	7	9	12	15	22

The probability of falsely rejecting the hypothesis $\sigma^2 = \sigma_0^2$ is equal to 0.05 throughout.

different from σ_1 .

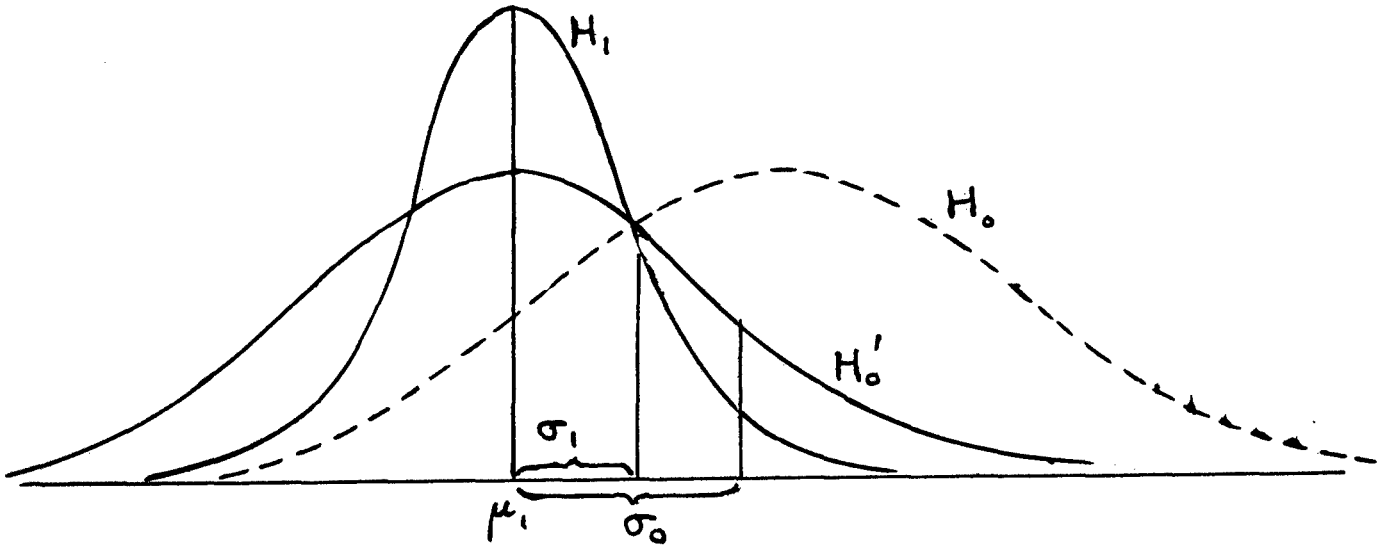


Figure 10. Distributions of X_i under hypotheses H_o, H_1, H_o'

For this case, then, we take H_o' as the simple hypothesis $\sigma = \sigma_o$ and $\mu = \mu_1$ (see Fig. 10) and the problem is reduced to that of § 3.2 with $\sum_i (x_i - \mu_1)^2$ instead of $\sum x_i^2$.

The test is $\sum (x_i - \mu_1)^2 < c$ (3.9)

where c is given by

$$\chi^2_{N,1-\alpha} = c / \sigma_o^2 \,. \tag{3.10}$$

Now, on the assumption of normality in the parent population, it may be easily shown that the probability that $\sum (x_i - \mu_1)^2 < c$ under H_o is never greater than under H_o' , whatever the value of μ . It follows, then, that (3.9) is a most powerful test (of size α) of H_o against H_1. Since this test depends on the value of μ_1, it is not uniformly most powerful against a composite alternative $\sigma = \sigma_1$, with μ unspecified.

When $\sigma_1 > \sigma_o$, the considerations given above do not apply. A concentrated distribution of H_o at $\mu = \mu_1$ would be <u>easier</u>

to distinguish from H_1 than any other. Now it has been shown by Neyman and Pearson that the test "reject H_o in favour of H_1 when $Nv > c$" , where v is the sample variance $\sum (x_i - \bar{x})^2 / N$, has the property of similarity, that is to say, the test has the same size whatever the value of μ , and this particular test is in fact most powerful among all similar tests for H_o against H_1. Lehmann and Stein, 1948, applied the method of § 2.5 to the case $\sigma_1 > \sigma_o$ by choosing as the distribution of μ that unique distribution which reduces the likelihood ratio test for H_o' against H_1 to the known most powerful similar test. In this case H_o' is the hypothesis that $\sigma^2 = \sigma_o^2$ and that the probability density $f(\mu)$ for any given μ is

$$f(\mu) = C e^{-\frac{N(\mu-\mu_1)^2}{2(\sigma_1^2-\sigma_o^2)}} \qquad (3.11)$$

The test of H_o' against H_1 is then $\sum (x_i - \bar{x})^2 > c$, c being given by

$$\Pr\left\{ \sum_i (x_i - \bar{x})^2 > c \;\middle|\; f(\mu), \sigma_o^2 \right\} = \alpha . \qquad (3.12)$$

This probability, however, is independent of the value of μ and depends only on σ_o and N. In fact,

$$\alpha = \int_{c/\sigma_o^2}^{\infty} f_{N-1}(\chi^2)\, d\chi^2 \qquad (3.13)$$

where $f_{N-1}(\chi^2)$ is the probability density for χ^2 with N - 1 degrees of freedom.

The test is therefore most powerful for H_o against H_1 and since it does not depend on μ_1 or σ_1 , it is a uniformly most powerful test against all alternatives.

This test can be extended to cover the cases

$$\begin{cases} H_o : \sigma \leq \sigma_o \\ H_1 : \sigma = \sigma_1 \; (> \sigma_o), \; \mu = \mu_1 \end{cases},$$

and

$$\begin{cases} H_0 : \sigma \geq \sigma_0 \\ H_1 : \sigma = \sigma_1 \ (< \sigma_0), \ \mu = \mu_1 \ . \end{cases}$$

In both of these, the least favourable distribution of σ would be a concentration at $\sigma = \sigma_0$. For the first case, it is clear from (3.13) that the size of the test $\leq \alpha$, for $\sigma \leq \sigma_0$ since c/σ^2 will then be equal to or greater than c/σ_0^2. For the second case, the same thing is true, since from (3.10)

$$\alpha = \int_0^{c/\sigma_0^2} f_N(\chi^2) \, d\chi^2 \qquad (3.14)$$

and the integral is reduced in value by putting σ^2 in place of σ_0^2 when $\sigma > \sigma_0$.

3.4 Power of the Above Test

Case 1: $\sigma_1 < \sigma_0$. The power of the test for H_0', and therefore for H_0, against H_1 is

$$P = \int_0^{c/\sigma_1^2} f_N(\chi^2) \, d\chi^2 , \qquad (3.15)$$

which is readily calculated from tables of the χ^2 distribution. Thus if $\alpha = 0.05$, $N = 5$, $\mu_1 = 0$, $\sigma_0 = 1$ and $\sigma_1 = 0.8$, we find from (3.10) that $c = 1.145$. Then from (3.15), $P = 0.125$. This is slightly greater than the value (0.110) given by using the sample variance.

The power of the test for H_0 against the composite alternative H_1 ($\sigma = \sigma_1$, μ unspecified) is

$$P = \Pr\left\{ \sum (x_i - \mu_1)^2 < c \mid \sigma_1, \mu \right\} \qquad (3.16)$$

c being still given by (3.10). We will take μ_1 as zero, as before.

The quantity $\sum x_i^2/\sigma_1^2$ now follows a non-central χ^2 distribution with N degrees of freedom, depending on the parameter

$$\lambda = N\mu^2/\sigma_1^2 \qquad (3.17)$$

The probability density is as given in (2.16), where $x = \sum x_i^2 / \sigma_1^2$. The power of the test is

$$P = \int_0^{c/\sigma_1^2} f(x)dx \qquad (3.18)$$

Numerical tables of the non-central χ^2 distribution, prepared by E. Fix, do not help us here, since, for one thing, they refer to the upper tail of the distribution, and, for another, they require that $\sigma_0 = \sigma_1$. The power may be expressed in the form

$$P = b^{-1} \int_0^{\infty} \phi\left(\frac{a-y}{b}\right) \cdot I\left(\frac{y}{(2N)^{1/2}}, \frac{N-2}{2}\right) dy, \qquad (3.19)$$

where $a = c/\sigma_1^2 - \lambda$, $b = 2\lambda^{1/2}$ and $I(u, p)$ is the Incomplete Gamma function tabulated by K. Pearson (1922).

The integrand in (3.19) vanishes at both limits and the integral may be approximately evaluated by quadrature.

For the numerical values mentioned above and for $\mu = 0.05$ 0.05, we have $\lambda/2 = 0.977$, $c/\sigma_1^2 = 1.789$, and the power P is about 0.095.

Case 2: $\sigma_1 > \sigma_0$. The power of the test for H_0 against H_1 is

$$P = \int_{c/\sigma_1^2}^{\infty} f_{N-1}(\chi^2)\, d\chi^2, \qquad (3.20)$$

where c is given by (3.13). The power is therefore independent of μ_1. Power curves derived from (3.16) for $\sigma_1 < \sigma_0$ and from (3.20) for $\sigma_1 > \sigma_0$, for a few values of N, are given in Fig. 9. If $\sigma_1 < \sigma_0$, the N given in this figure should be reduced by 1.

3.5 Composite Hypothesis against Composite Alternative

A more general case of testing arises when both the null hypothesis and the alternative hypothesis are composite. Symbolically,

$$\begin{cases} H_0 : \theta \in \omega \\ H_1 : \theta \in \Omega - \omega \end{cases}$$

where ω is a specified region of the whole space Ω available for θ. If $\psi(X)$ is a test of size α, we want it to be such that

$$\left.\begin{aligned} &\text{(i)} \int \psi(x)\, dF(x|\theta) \leq \alpha \quad \text{for} \quad \theta \in \omega, \\ &\text{(ii)} \int \psi(x)\, dF(x|\theta) \quad = \max, \text{ for } \theta \in \Omega - \omega. \end{aligned}\right\} \tag{3.21}$$

These conditions mean that (i) the probability of rejecting the null hypothesis, if true, is not greater than α and (ii) the probability of rejecting it if false is a maximum. If a test that satisfies (i) also satisfies (ii) for all values of θ, it is U.M.P., but usually this is not so. Sometimes, however, we can get a U.M.P. test if we restrict ourselves to a particular class of tests which possess some desirable property. Among such properties are those of invariance and unbiasedness. The meaning of an unbiased test has been discussed in § 1.10. An invariant test is one which is invariant under some suitable transformation of variables in the sample space - naturally we will choose some simple and obvious transformation such as a translation or a change of scale.

IV. STUDENTS' t -TEST

4.1 The one-tailed t-test

In the examples given in Chapter II the variance under the null hypothesis was supposed known. In 1908, W.S. Gosset ("Student") introduced the now familiar test for the mean of a normal population, a test which depends on the sample mean and the sample variance but which is independent of the population variance.

Suppose that a sample of size N is taken from a normal population of mean μ and variance σ^2. Let the null hypothesis be $H_0 : \mu = \mu_0$, σ unspecified.

Without loss of generality we can put $\mu_0 = 0$ (we merely need to subtract μ_0 from all the observed sample values.) An alternative simple hypothesis H_1 is that $\mu = \mu_1$, and $\sigma = \sigma_1$, where we will suppose first that $\mu_1 > 0$.

Under the null hypothesis, the statistic $T = \bar{x}\left[\frac{N(N-1)}{\Sigma(x_i - \bar{x})^2}\right]^{1/2}$ has the Student t-distribution with N - 1 degrees of freedom. The probability density is

$$f(t) = A\left[1 + t^2/(N-1)\right]^{-N/2} \qquad (4.1)$$

where $1/A = (N-1)^{1/2} * B\left[(N-1)/2,\ 1/2\right]$ and $B[a,b]$ is the Beta function of a and b. Therefore f(t) is independent of σ The distribution is symmetrical about 0, which is the expected value of T. If the observed value of T exceeds t_α, where t_α is so chosen that

$$\int_{t_\alpha}^{\infty} f(t)\, dt = \alpha, \qquad (4.2)$$

and if we agree to reject H_0 when $T > t_\alpha$, then we shall clearly commit an error of the first kind with probability α. Since we are considering only alternatives with $\mu_1 > 0$, we are using only the upper tail of the t-distribution, and the test is one-tailed.

If we suppose that $\mu_1 < 0$, we shall reject H_o when $-T > t_\alpha$, and this is also a one-tailed test, using the lower tail.

If the null hypothesis is $H_o : \mu \leq 0$ $(\mu_1 > 0)$ the probability of error of the first kind will not be greater than α. This follows because the t-distribution is symmetrical with a maximum at $t = 0$.

4.2 The t-test and maximum likelihood

The probability density for a particular observed set of sample values $x_1, x_2, \ldots, x_n$, on the null hypothesis, is

$$p_o = (2\pi\sigma^2)^{-N/2} e^{-\Sigma x_i^2 / 2\sigma^2}$$

so that

$$L_o = \log p_o = -\frac{N}{2} \log (2\pi\sigma^2) - \Sigma x_i^2 / 2\sigma^2$$

$$= C - N \log \sigma - \frac{N}{2\sigma^2} (\bar{x}^2 + s^2) \qquad (4.3)$$

where $\bar{x}$ and s^2 are the sample mean and sample variance respectively. On the alternative hypothesis the probability density is p_1, and

$$L_1 = \log p_1 = C - N \log \sigma_1 - \frac{N}{2\sigma_1^2} \left[(\bar{x} - \mu_1)^2 + s^2 \right] \qquad (4.4)$$

From (4.3), L_o (and therefore p_o) is constant over the surface $\Sigma x_i^2 =$ constant. Suppose we pick a region of rejection on each such surface, equal in area to a fraction α of the area of the surface. If $\psi = 1$ when x lies in this region and 0 otherwise, the expectation of ψ for a given value of Σx_i^2 will be α. The whole region of rejection R will be a combination of the regions for all possible values of Σx_i^2, and it is obviously independent of σ. We have

$$\int_{(R)} p_o \, dx_1 \, dx_2 \ldots dx_N = \alpha . \qquad (4.5)$$

The test will be most powerful if the probability of rejection of H_o

when H_1 is true is as great as possible. That is, we must choose R so as to maximize

$$P = \int_{(R)} p_1 \, dx_1 \, dx_2 \ldots dx_N \, , \tag{4.6}$$

subject to the condition (4.5). The solution of this problem for $\mu_1 > 0$ by the method of Lagrange multipliers (see, for example, Kenney and Keeping, 1951, pp. 392-3), leads to the conclusion that R is defined by the condition $t > t_\alpha$, where t_α is given by (4.2) and t has the distribution (4.1). The method of maximum likelihood therefore leads to the ordinary one-tailed t-test, and since this test is independent of the particular value of μ_1 chosen, it is uniformly most powerful (U. M. P.) against any H_1 with $\mu_1 > 0$, whatever the values of σ and σ_1 may be. Similarly, if $\mu_1 < 0$ the test - $t > t_\alpha$ is U. M. P., but if μ_1 ~~must~~ may be greater or less than zero no U. M. P. test exists.

4.3 The power of the t-test

The power of the test is the probability of rejecting H_0 when H_1 is true, and is therefore given by (4.4) and (4.6) with R defined by $t > t_\alpha$. Now the probability of a particular set of observed values under H_1, namely $p_1 \, dx_1 \ldots\ldots dx_N$, can be expressed as $f(\bar{x}, s) \, d\bar{x} \, ds$ where $f(\bar{x}, s)$ is the joint probability density for x and s, and is given by

$$f(\bar{x}, s) = K \, s^{N-2} \, e^{-\frac{N}{2\sigma_1^2}\left[(\bar{x}-\mu_1)^2 + s^2\right]} \tag{4.7}$$

where $K = 2\pi^{-1/2} (N/2)^{-N/2} \sigma_1^{-N} / \Gamma\left(\frac{N-1}{2}\right)$.

Since $t = (N-1)^{\frac{1}{2}} \bar{x}/s$, we can find P by integrating (4.7) over all $\bar{x}$ such that $\bar{x} > (N-1)^{-1/2} s \, t_\alpha$ and over all s from 0 to ∞. (For any point in this region, $t > t_\alpha$.) That is,

$$P = K \int_0^\infty s^{N-2} e^{-Ns^2/2\sigma_1^2} \int_{(N-1)^{-1/2} s \, t_\alpha}^\infty e^{-N(\bar{x}-\mu_1)^2/2\sigma_1^2} \, d\bar{x} \, ds$$

On putting $z = N^{1/2}(\bar{x} - \mu_1)/\sigma_1$ and $\chi^2 = Ns^2/\sigma_1^2$, so that z is a standard normal variate, and χ^2 has the χ^2 distribution with N - 1 degrees of freedom, we find

$$P = \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{\frac{n-2}{2}}} \int_0^\infty \chi^{n-1} e^{-\chi^2/2} \int_{t_\alpha \chi n^{-1/2} - \rho_1}^\infty (2\pi)^{-1/2} e^{-z^2/2} \, dz \, d\chi$$

where n = N - 1, and $\rho_1 = N^{1/2}\mu_1/\sigma_1$. This may be expressed, by a change of variable, as

$$P = \int_0^\infty f_n(\chi^2) \left[1 - \Phi\left(t_\alpha \chi n^{-1/2} - \rho_1\right)\right] d\chi^2. \qquad (4.8)$$

The integral can be evaluated numerically when α, n and ρ_1 are given, so that the power of the test is obtainable as a function of μ_1 and σ_1. The power function is not independent of σ_1; Dantzig (1940) showed that no test of the hypothesis H_o can exist with a power function independent of σ_1. In a number of practical cases, however, we do have some rough knowledge of σ_1, and if so we can use the tables calculated from (4.8) for estimating the size of sample necessary to detect a difference between μ and o of a given order of magnitude.

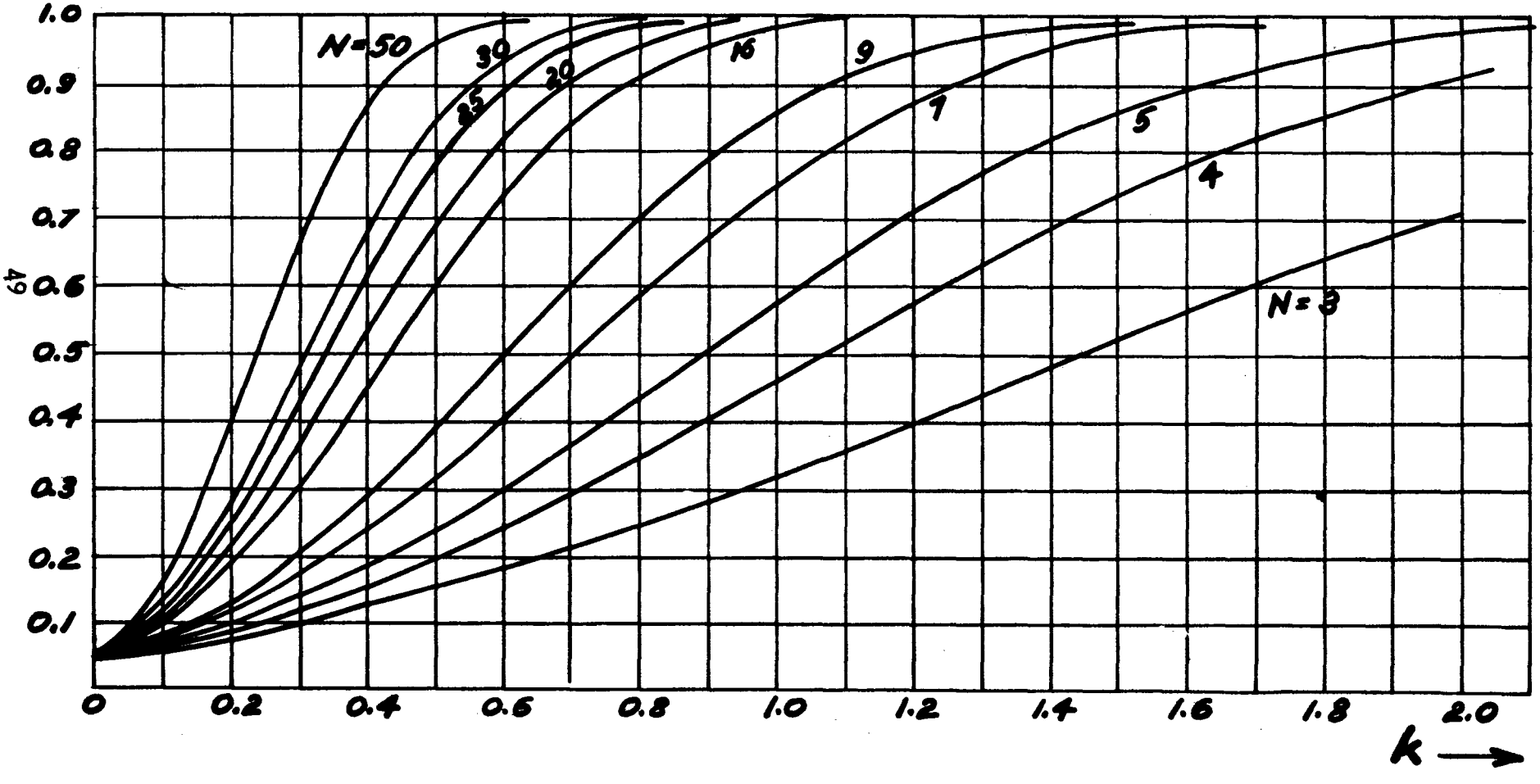
As an example of the use of this method, suppose a new treatment is under investigation, intended to increase the strength of a certain alloy. The claim that it does produce an increase is tested on a sample of N pairs, each pair consisting of one specimen of the alloy having undergone the new treatment and one specimen having had the standard treatment, the members of a pair being in other respects as alike as possible. The increase in strength is measured for each pair. We shall not be interested in a possible decrease in strength, and so we shall use a one-tailed test. The standard deviation of many measured strengths under the standard treatment may be taken as an estimate of σ, and the size of sample necessary to detect an average increase of strength equal to k σ can be determined for fixed values of the probabilities of error of the first and second kind.

4.4. Tables of the Power Function

If t is defined as $n^{1/2}\bar{x}/s = n^{1/2}(z + \rho_1)/\chi$, then, on the null

FIG. 11. POWER CURVES for the ONE-TAILED T- TEST ($\alpha = 0.05$)

$$(k = \mu_1/\sigma_1)$$



hypothesis H_0 ($\rho_1 = 0$), t has the ordinary Student t-distribution, but when $\mu = \mu_1$ and $\sigma = \sigma_1$, (that is, on the alternative hypothesis H_1) t has what is known as the non-central t-distribution. The probability density of t is

$$f(t) = A\left(1 + t^2/n\right)^{-(n+1)/2} e^{-\frac{\rho_1^2}{2(1+t^2/n)}} H\left(-\frac{t\rho_1}{(t^2+n)^{1/2}}\right) \qquad (4.9)$$

where $1/A = n^{1/2}\, B\left[n/2,\ 1/2\right]$

and

$$H(x) = \frac{1}{2^{(n-1)/2}\,\Gamma\left(\frac{n+1}{2}\right)} \int_0^\infty v^n e^{-(v+x)^2/2}\, dv .$$

When $\rho_1 = 0$, $H(0) = 1$ and (4.9) reduces to the form (4.1). The probability integral of f(t) is the power of the t test,

i.e. $$P = \int_{t_\alpha}^{\infty} f(t)\, dt ,$$

which can be shown to be equivalent to (4.8).

Tables of P were calculated by Johnson and Welch (1939). These tables are necessarily of triple entry, since P depends on n, ρ_1 and α. They may be arranged in various ways; for example, to give ρ_1, for selected values of n, α and P ($= 1 - \beta$), or to give α for selected values of ρ_1, n and P. The former arrangement was preferred by Johnson and Welch.

Suppose that α and n are given. Then t_α is determined from the ordinary t - table, and ρ_1 can be found, after some preliminary calculations, from the tables of Johnson and Welch. Directions for the use of the tables are given on p. 272 of the reference cited.

Shorter tables, which however are better adapted to the particular problem of the power of the one-tailed t-test, were calculated by Neyman and Tokarska (1936). These give for α = 0.05 and 0.01, and for all n from 1 to 30 the values of ρ_1 corresponding to selected values of β. The curves in Fig. 11 were drawn from these tables. They show for selected values of N the power P corresponding to $k = \mu_1/\sigma_1 = \rho_1 N^{-1/2}$

Suppose for instance we know that $\sigma_1 = 10$ and N = 17. If the chances of each kind of error are equal to 0.05, we find from the Neyman and Tokarska tables that $\rho_1 = 3.44$, so that $\mu_1 = N^{-1/2}\sigma_1\rho_1 = 8.34$. This means that we have a 95% chance of detecting a real difference in the mean equal to 0.83 of the standard deviation, when the probability of apparently detecting such a difference when none really exists is 0.05. The result can be roughly checked by Fig. 11.

As a rough approximation for moderate-sized samples,

$$\rho_1 \approx t_\alpha - \chi_p \left(1 + t_\alpha^2/2n\right)^{1/2} \qquad (4.10)$$

where χ_p is the standard normal variate which is exceeded with probability P.
With the data above, $\chi_p = -1.645$, so that $\rho_1 = 1.746 + 1.645\,(1.095)^{1/2}$
$= 3.47$, which is not greatly in error.

Another approximation, for $N > 10$, is

$$P = \Pr\left\{ t < \rho_1 - t_\alpha \right\} \qquad (4.11)$$

where t is the ordinary central t with N - 1 degrees of freedom.

Thus, for P = 0.95, with N = 17, $\rho_1 = t_\alpha + 1.746 = 3.49$.

4.5 Test of the Difference of Means in two Samples

Suppose two samples of sizes N_1 and N_2 give means of $\bar{x}_1$ and $\bar{x}_2$, the true population means being μ_1 and μ_2, and the standard deviation σ being the same for both populations. The null hypothesis H_o is that $\mu_2 - \mu_1 \leq 0$ and the alternative hypothesis H_1 is that $\mu_2 - \mu_1 = k\sigma\ (k > 0)$. The statistic $\dfrac{\bar{x}_2 - \bar{x}_1}{\hat{\sigma}}\left[\dfrac{1}{N_1} + \dfrac{1}{N_2}\right]^{-1/2}$ where $\hat{\sigma}^2$ is an estimate of σ^2 equal to $(N_1 s_1^2 + N_2 s_2^2)/(N_1 + N_2 - 2)$,

has the t distribution with $N_1 + N_2 - 2$ degrees of freedom, on the hypothesis that $\mu_2 - \mu_1 = 0$. We reject this hypothesis when $t > t_\alpha$, the chance of error in so doing being α. The chance of error in rejecting H_o (that $\mu_2 - \mu_1 \leq 0$) is therefore <u>not greater than</u> α.

The quantity corresponding to ρ_1 in the above theory is now

$\rho = \frac{\mu_2 - \mu_1}{\sigma} \left(\frac{N_1 N_2}{N_1 + N_2} \right)^{1/2}$, and the number of degrees of freedom is $n = N_1 + N_2 - 2$.

Suppose, for example, that $N_1 = N_2 = 10$, so that $n = 18$. The value of ρ for $\alpha = 0.05$ and $P = 0.9$ is 3.04, and therefore $(\mu_2 - \mu_1)/\sigma = 1.36$. This indicates that a difference in the means of the two samples as great as $1.36\,\sigma$ would stand a chance of 0.9 of being detected by the proposed test.

This result could be approximately read from Fig. 11, but in using this figure for the two-sample problem we must take N as n + 1 (in this case 19) and multiply the value of k by $(n+1)^{1/2} \left(\frac{N_1 N_2}{N_1 + N_2} \right)^{-1/2} = 1.95$. The value read from the figure is about 0.7, which gives $(\mu_2 - \mu_1)/\sigma$ = 1.4.

As another example, let us suppose that we are interested in the difference of tensile strength between two types of casting, the variability being about the same in the two types, and that we use a t-test on samples of N of each type. For a given $\alpha = 0.05$, say, we can calculate the power corresponding to an assigned k, that is, the chance of detecting a superiority of the one type over the other equal in amount to $k\,\sigma$. We now have $\rho = k\,[N/2]^{1/2}$ and $n = 2N - 2$. Table V, calculated from Neyman and Tokarska's tables, gives k for certain values of N. It shows, for instance, that if we want to have at least an even chance of detecting a superiority in the mean equal to one standard deviation, we should use samples of at least 6 or 7 each.

4.6 The two-tailed t-test

In many problems we are more concerned with the magnitude of the difference between the true population mean μ and some assumed value μ_0 than with its sign. We may want to be reasonably sure, for example, that the mean thickness of a batch of mica washers does not differ by more than a set amount from its nominal value, but not be particularly concerned over whether the washers happen to be a little too thick rather than too thin. The null hypo-

thesis is $\mu = \mu_0$ and the alternative hypothesis $\mu = \mu_1$, $\sigma = \sigma_1$, where $|\mu_1 - \mu_0| = k\,\sigma_1$. The null hypothesis is rejected when $|t| > t_\alpha$, t_α being determined by

$$\int_{-t_\alpha}^{t_\alpha} f_0(t)\,dt = 1 - \alpha, \tag{4.12}$$

where $f_0(t)$ is the probability density for central t.

The power is given by

$$P(\mu_1, \sigma_1) = 1 - \int_0^\infty f_n(\chi^2)\left[\Phi\left(t_\alpha \chi n^{-1/2} - \rho_1\right) - \Phi\left(-t_\alpha \chi n^{-\frac{1}{2}} - \rho_1\right)\right] d\chi^2 \tag{4.13}$$

where $\rho_1 = N^{\frac{1}{2}}(\mu_1 - \mu_0)/\sigma_1 = \pm k N^{\frac{1}{2}}$, $n = N - 1$.

The non-central t distribution is skew, and if k is large the area in one tail is negligible. Unless N is very small this is true for quite moderate values of k. Thus for N = 10 and k = 0.216, the probability that $t < -t_\alpha$ (if $\mu_1 > \mu_0$) or that $t > t_\alpha$ (if $\mu_1 < \mu_0$) is less than 0.005. In the former case, the power is practically equal to $\int_{t_\alpha}^\infty f_1(t)\,dt$, where $f_1(t)$ is the probability density for non-central t. But by (4.12),

$$\int_{t_\alpha}^\infty f_0(t)\,dt = \alpha/2, \tag{4.14}$$

so that we can use the tables of the one-tailed test for the present problem, provided we remember that when these tables specify $\alpha = 0.05$, we are really using $\alpha = 0.10$.

TABLE V

Power of the t-test for distinguishing between the Means of two Samples of size N, with common σ. (Values of k such that P is the probability of detecting a difference equal to k σ, when $\alpha = 0.05$.)

N \ P	0.2	0.4	0.5	0.6	0.8	0.9	0.95
3	0.78	1.36	1.62	1.88	2.48	2.94	3.32
4	0.64	1.11	1.32	1.52	1.99	2.35	2.65
5	0.56	0.96	1.14	1.32	1.73	2.03	2.29
6	0.50	0.86	1.02	1.18	1.54	1.82	2.04
8	0.42	0.73	0.86	1.00	1.31	1.54	1.73
10	0.37	0.65	0.76	0.88	1.16	1.36	1.53
12	0.34	0.59	0.69	0.80	1.05	1.23	1.39
14	0.31	0.54	0.64	0.74	0.96	1.14	1.28
16	0.29	0.50	0.59	0.69	0.90	1.06	1.19
25	0.23	0.40	0.47	0.54	0.71	0.84	0.94
50	0.16	0.28	0.33	0.38	0.50	0.59	0.67

4.7 The One-tailed Test as an Invariant Test

We arrive at the same test by searching for the most powerful invariant test of the hypothesis $H_o : \mu \leq 0$ against $H_1: \mu > 0$, the variance of the parent population being unknown. The mean M and the variance V jointly form a sufficient statistic. If we put $T = M/V^{1/2}$, then T is invariant for the class of transformations $X_i' = c X_i$, where c is a positive constant, since of course $M' = cM$ and $V' = c^2V$. (This transformation merely amounts to a change of scale in the measurement of X).

The T so defined differs from the T defined in § 4.1 only by not having the constant factor $(N - 1)^{1/2}$, and so is essentially the same.

If $Z = M/\sigma$ and $W = V/\sigma^2$, then on the hypothesis H_1, Z is normal with mean $\delta = \mu/\sigma$, and W has the χ^2 distribution with N - 1 degrees of freedom. The joint probability that Z lies between z and z + dz and W between w and w + dw is therefore

$$C e^{-N(z-\delta)^2/2} w^{(N-3)/2} e^{-w/2} dz\, dw.$$

Now $T = Z/W^{1/2}$, so that for a given value w of W, $T = w^{-1/2} Z$. If $f_\delta(t, w)\, dt\, dw$ is the joint probability for t and w, with $dt = w^{-1/2} dz$, we have

$$f_\delta(t, w) = C e^{-N(tw^{1/2}-\delta)^2/2} w^{(N-2)/2} e^{-w/2} \tag{4.15}$$

The probability density for t, regardless of the value of w, is given by integrating (4.15) over w from 0 to ∞. That is,

$$f_\delta(t) = C \int_0^\infty e^{-N(tw^{1/2}-\delta)^2/2} e^{-w/2} w^{(N-2)/2} dw \tag{4.16}$$

The null hypothesis is equivalent to $\delta \leq 0$, and the alternative hypothesis to $\delta > 0$. If we choose a particular alternative $\delta_1 > 0$, and apply the method of § 2.6, the difficulty of distinguishing between H_o and H_1 may be expected to be as great as possible when $\delta = 0$. If we let H_o' be the hypothesis that $\delta = 0$, the most powerful invariant test of size α for distinguishing between H_o' and H_1 will be

$\psi(t) = 1$ when $f_0(t)/f_\delta(t) < c$.

This ratio can be written

$$F(t) = \left[\int_0^\infty f(v,t)\,dv\right] \Big/ \left[k\int_0^\infty e^{N\delta_1 v} f(v,t)\,dv\right] \qquad (4.17)$$

where (for $t > 0$) $v = w^{1/2}t$, $k = e^{-N\delta_1^2/2}$ and $f(v,t) = v^{N-1}e^{-Nv^2/2}e^{-v^2/2t^2}$,

and it is not difficult to show that $F(t)$ is a strictly decreasing function of t, i.e. $(F(t_1) < F(t_2)$ if and only if $t_1 > t_2$. This means that the test can be written

$\psi(t) = 1$ when $t > c$,

where c is now determined by

$$\Pr\{t > c \mid H_0'\} = \alpha, \qquad (4.18)$$

and so is given by the ordinary table of t.

The same conclusion holds for $t < 0$, v in (4.17) being now equal to $-w^{1/2}t$, and $e^{N\delta_1 v}$ replaced by $e^{-N\delta_1 v}$.

Now, from the shape of the t distribution, it follows that the probability that $t > c$, under $H_0: \mu \leq 0$, is never greater than it is under H_0' : $\mu = 0$ (compare Fig. 7, which is drawn for the normal curve with mean μ_0. The general shape of the t-distribution resembles this curve). That is,

$$\Pr\{t > c \mid H_0\} \leq \alpha.$$

It follows that the test is most powerful for distinguishing H_1 from H_0. Since it is independent of the particular value δ_1 chosen, it is U. M. P. among invariant tests for distinguishing H_0 ($\mu \leq 0$) from H_1 ($\mu > 0$).

4.8 The Two-tailed Test as Invariant Test

Here we want to distinguish between H_0 ($\mu = 0$) and H_1 ($\mu \neq 0$), or, in terms of the quantity $\delta = \mu/\sigma$ introduced in § 4.7, between

$\delta = 0$ and $\delta \neq 0$. If we again use the Lehmer and Stein method, of replacing H by a "most difficult" simple hypothesis H_1', it would seem reasonable to take for H_1' the hypothesis that δ is equally like to be $\pm \delta_1$, where δ_1 is a fixed number greater than 0. Instead of (4.16) we have

$$f_\delta(t) = \frac{C}{2} \int_0^\infty \left[e^{-\frac{N}{2}(tw^{\frac{1}{2}} + \delta_1)^2} + e^{-\frac{N}{2}(tw^{\frac{1}{2}} - \delta_1)^2} \right] e^{-\frac{w}{2}} w^{\frac{N-2}{2}} dw \tag{4.19}$$

and instead of (4.17)

$$F(t) = \left[\int_0^\infty f(v, t)\, dv \right] \Big/ \left[k \int_0^\infty f(v, t) \cosh(N\delta_1 v)\, dv \right] \tag{4.20}$$

It can be shown that F(t) decreases as t increases for $t > 0$ and decreases as t decreases when $t < 0$. The test $\psi(t) = 1$ when $F(t) < c$ is therefore equivalent to $\psi(t) = 1$ when $|t| > c$, and this is the ordinary two-sided Student test, c being given by

$$\Pr\{ |t| > c \mid H_0 \} = \alpha .$$

Since it is independent of δ_1 this test is U. M. P. for H_0 against the composite alternative H_1 (namely, $|\delta| > 0$).

TESTS FOR THE PROPORTION DEFECTIVE

5.1 Simple Hypothesis against Simple Alternative

Let p be the observed proportion of items in a sample of N which have a certain characteristic. For convenience we shall refer to this characteristic as that of being "defective" but of course it may be of quite a different nature. It is merely necessary that an inspector shall be able to say unambiguously of each item whether or not it possesses the characteristic in question (which might for instance be that of being of the male sex if the sample consists of animals).

If π is the true proportion defective in the population (assumed to be very large) and X is the observed number defective out of a sample of N (X is a random variable), then

$$\Pr\{X = x\} = \binom{N}{x} \pi^{x} (1 - \pi)^{N-x} \qquad (5.1)$$

where $\binom{N}{x} = N! \,/\, [x!\ (N - x)!]$

Let the null hypothesis H_0 be that $\pi = \pi_0$ and the alternative hypothesis H_1 that $\pi = \pi_1$. Then in the notation of § 1.~~17~~.7

$$f(x \mid \pi_0) = \binom{N}{x} \pi_0^{x} (1 - \pi_0)^{N-x} = f(x),$$

$$f(x \mid \pi_1) = \binom{N}{x} \pi_1^{x} (1 - \pi_1)^{N-x} = g(x),$$

and $L(x) = x \log (\pi_0/\pi_1) + (N - x) \log \{(1 - \pi_0)/(1 - \pi_1)\}$.

The condition for rejection, $L(x) < c$, is therefore

$$x \left[\log \frac{\pi_0}{\pi_1} + \log \frac{1 - \pi_1}{1 - \pi_0} \right] + N \log \frac{1 - \pi_0}{1 - \pi_1} < c . \qquad (5.2)$$

First let us suppose that $\pi_1 > \pi_0$, so that $1 - \pi_1 < 1 - \pi_0$. The coefficient of x in (5.2) is negative, and (5.2) is equivalent to

$$x > c \qquad (5.3)$$

where c is a new constant. The value of c is determined by the size of the test, i.e. by

$$\sum_{x=0}^{N} \psi(x)\, f(x) \leq \alpha . \tag{5.4}$$

As described in § 1.8, we can choose a suitable integer c, and a suitable probability ψ_0, so that for the test

$$\begin{cases} \psi(x) = 1, & x > c, \\ \psi(x) = 0, & x < c, \\ \psi(x) = \psi_0, & x = c, \end{cases}$$

(which means that we reject H_0 if $x > c$, and reject it with probability ψ_0 when $x = c$) we have

$$f(c)\, \psi_0 + \sum_{c+1}^{N} f(x) = \alpha . \tag{5.5}$$

The power of the test is

$$P = g(c) \cdot \psi_0 + \sum_{c+1}^{N} g(x) . \tag{5.6}$$

For given values of α, π_0 and N (≤ 100) we can use the Tables of the Binomial Probability Distribution (National Bureau of Standards, 1950, and H. G. Romig, 1953) to find P as a function of π_1, and hence construct power curves for this test. For large N and π_0 near 0.5 the distribution of x under hypothesis H_0 is approximately normal with mean $N\pi_0$ and variance $N\pi_0 (1 - \pi_0)$. For large N and π_0 near 0 the distribution is approximately of the Poisson type, with parameter $N\pi_0$.

If $\pi_1 < \pi_0$, the test (5.2) is equivalent to $x < c$, where c is determined by

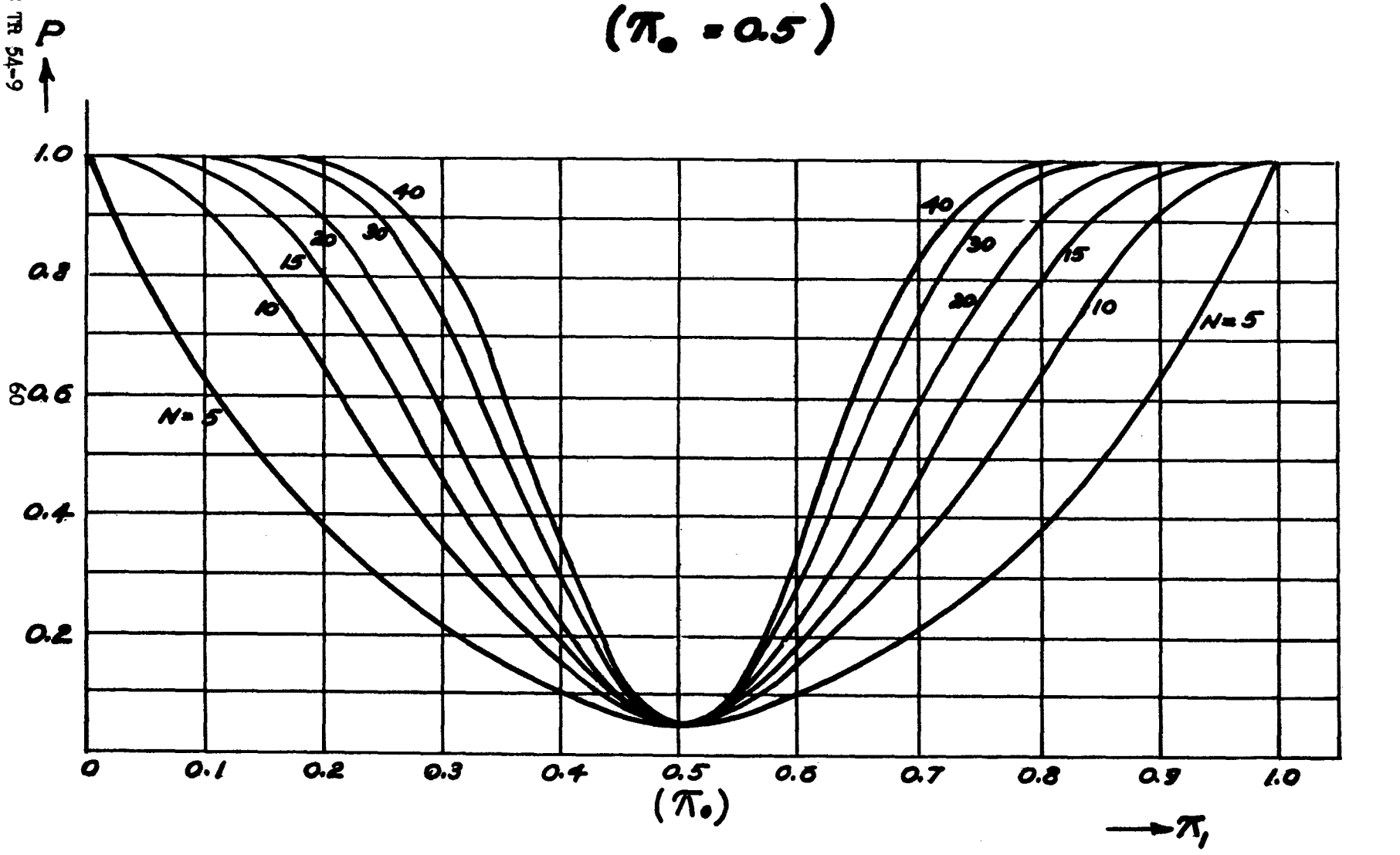
$$f(c)\, \psi_0 + \sum_{0}^{c-1} f(x) = \alpha , \tag{5.7}$$

which is equivalent to

$$\sum_{c}^{N} f(x) - f(c)\, \psi_0 = 1 - \alpha . \tag{5.8}$$

FIG. 12. POWER CURVES for the TEST of PROPORTION DEFECTIVE

($\pi_0 = 0.5$)



The power is now given by

$$1 - P = \sum_{c}^{N} g(x) - g(c)\ y_0 \qquad (5.9)$$

Figure 12 gives the symmetrical power curves with $\pi_0 = 0.5$ for several values of N, and Figure 13 illustrates the non-symmetrical case for $\pi_0 = 0.2$. The data for these curves and for some other values of π_0 are included in Table VI.

It is apparent from Figure 12 that to have at least a 50% chance of detecting the difference between an actual proportion 0.7 in the population and an assumed value 0.5, one would need a sample of about 16. To raise this chance to 80% one would need a sample of nearly 40. This is on the assumption that the chance of falsely stating that such a difference from 0.5 exists is only 0.05.

5.2 Proportion Defective when Distribution is Normal

We suppose that the objects tested have a characteristic X which is normally distributed with mean μ and standard deviation σ, and that if X is greater than some fixed value x_0, the object is classed as defective. We might think of bolts, for instance, which must be less than a certain diameter to pass through a hole of fixed size. The proportion π of defectives in the population will be the area under a standard normal curve beyond the ordinate at $(x_0 - \mu) / \sigma$. If $\bar{x}$ and s are the mean and the standard deviation of X in a sample of N items from the population, an estimate of $N^{1/2} (x_0 - \mu) / \sigma$, which we will call ρ, is provided by the statistic $u = n^{1/2} (x_0 - \bar{x}) / s$, n being written for N - 1. The probability p that a standard normal variate exceeds $N^{1/2}u$, i.e. $1 - \Phi(N^{1/2}u)$, is then an estimate of π.

Now $u = n^{1/2} \left(\frac{x_0 - \mu}{\sigma} - \frac{\bar{x} - \mu}{\sigma} \right) / \frac{s}{\sigma}$

$$= n^{1/2} (\rho - z) / \chi_n \qquad (5.10)$$

where $z = N^{1/2} (\bar{x} - \mu) / \sigma$ and $\chi_n = N^{1/2} s / \sigma$.

FIG. 13. POWER CURVES for TEST of PROPORTION DEFECTIVE.

($\pi_o = 0.2$)

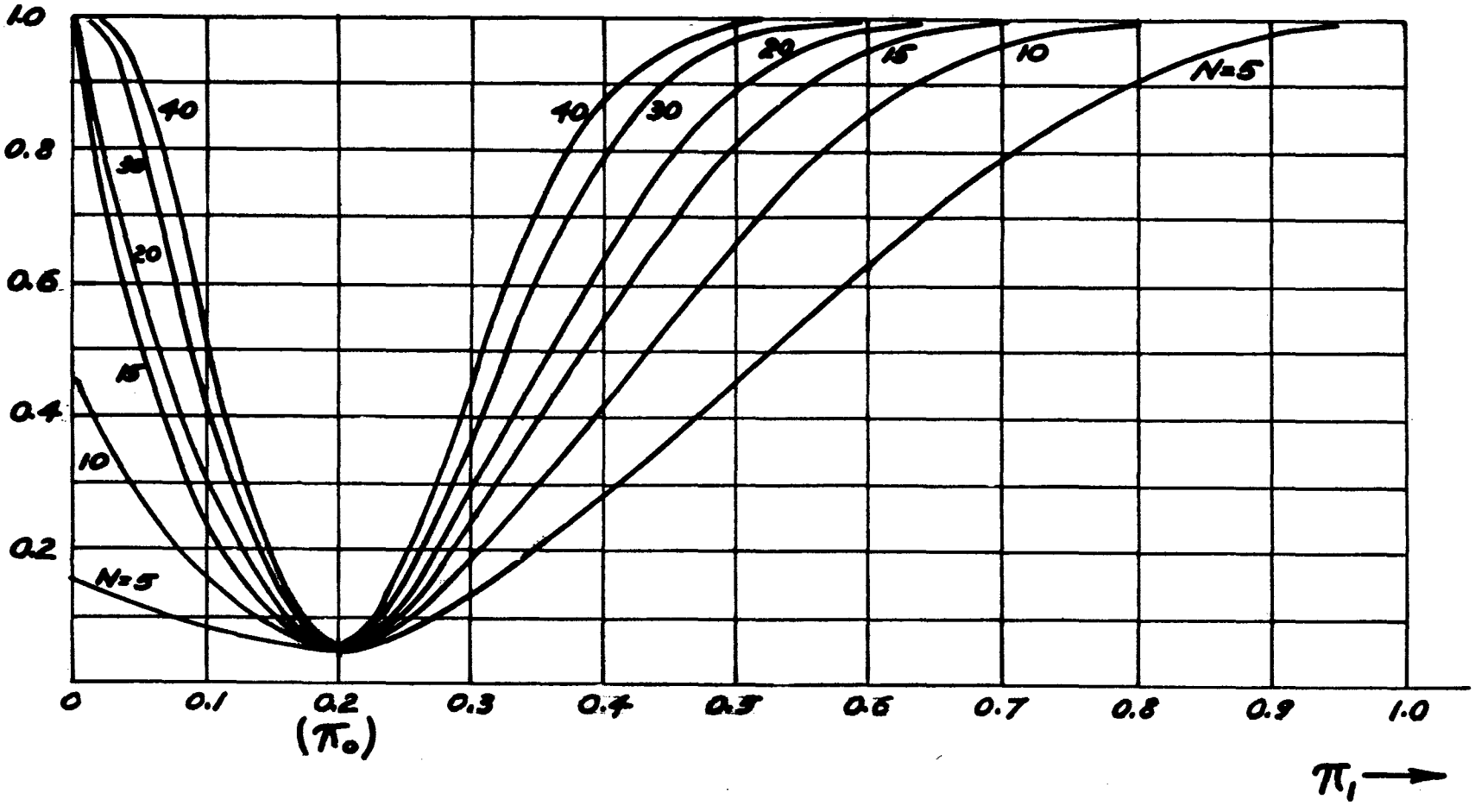


TABLE VI

Power of Test for Proportion Defective

This gives for assigned π_o, and $\alpha = 0.05$, the power of the randomized test for the proportion π_1 of defectives in the population against an assumed value π_o, for a sample of size N. The test consists in rejecting H_o if the observed number x of defectives is greater than c or less than c'. If $x = c$, H_o is rejected with probability ψ_o, and if $x = c'$ with probability ψ_o'.

(a) $\pi_o = 0.2$

N	c / c'	ψ_o / ψ_o'	π_1 .02	.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5	3 0	.845 .153	.138	.118	.090	.05	.143	.282	.452	.629	.789	.910	.980
10	4 0	.195 .466	.381	.279	.162	.05	.189	.416	.663	.856	.960	.995	1.000
15	6 1	.743 .112	.764	.504	.244	.05	.241	.544	.810	.950	.993	1.000	
20	7 1	.327 .667	.849	.610	.302	.05	.282	.638	.893	.984	.999		
30	10 3	.687 .074	.980	.822	.429	.05	.367	.788	.970	.999			
40	12 4	.153 .454	.995	.903	.517	.05	.444	.880	.992				

(b) $\pi_0 = 0.3$

N	c / c′	ψ_0 / ψ_0'	π_1 .05	0.1	0.2	0.25	0.3	0.35	0.4	0.5	0.6	0.7	0.8	0.9
5	3 0	.145 .298	.230	.176	.098	.071	.05	.080	.121	.233	.387	.573	.767	.929
10	5 1	.026 .180	.655	.418	.156	.090	.05	.099	.171	.383	.638	.852	.968	.998
15	7 2	.000 .161	.851	.592	.204	.105	.05	.113	.213	.500	.787	.950	.996	1.000
20	9 3	.031 .203	.937	.715	.248	.118	.05	.125	.250	.593	.875	.983	.999	
30	13 5	.224 .427	.990	.868	.329	.143	.05	.147	.316	.733	.958	.998		
40	17 7	.576 .832	.999	.948	.412	.168	.05	.169	.381	.832	.987	1.000		

(c) $\pi_o = 0.4$

N	c	ψ_o	π_1											
	c'	ψ_o'	.05	0.1	0.2	0.3	0.35	0.4	0.45	0.5	0.6	0.7	0.8	0.9
5	4	.518	.498	.380	.211	.108	.075	.05	.077	.112	.212	.355	.540	.760
	0	.643												
10	7	.888	.916	.742	.385	.156	.091	.05	.094	.159	.358	.620	.857	.981
	2	.030												
15	9	.264	.970	.840	.445	.159	.082	.05	.105	.191	.458	.761	.950	.998
	3	.187												
20	12	.816	.997	.954	.624	.234	.116	.05	.117	.230	.563	.782	.986	1.000
	4	.973												
30	16	.039	1.000	.993	.775	.301	.137	.05	.139	.298	.719	.961	.999	
	8	.128												
40	21	.308		.999	.870	.363	.156	.05	.158	.355	.815	.989		
	11	.414												

(d) $\pi_o = 0.5$

N	c	c′	$\psi_o = \psi_o'$	π_1				
				0.1 0.9	0.2 0.8	0.3 0.7	0.4 0.6	0.5
5	4	1	0.120	.630	.377	.211	.109	.05
10	8	2	0.893	.909	.646	.358	.154	.05
15	11	4	0.778	.978	.794	.467	.189	.05
20	14	6	0.793	.996	.891	.568	.224	.05
30	19	11	0.0124	1.000	.975	.732	.293	.05
40	25	15	0.264	1.000	.993	.828	.350	.05

For values of $\pi_o > 0.5$, use the above tables with $1 - \pi_o$ and $1 - \pi_1$ instead of π_o and π_1. The values of c and c′ will be N - c′ and N - c of the above tables respectively, and the values of ψ_o and ψ_o' will be interchanged. Thus for $\pi_o = 0.6$ and N = 40, c = 29 and c′ = 19, $\psi_o = 0.414$ and $\psi_o' = 0.308$.

In this expression, z is a standard normal variate and χ^2_n has the χ^2 distribution with n degrees of freedom; u, therefore, has the non-central t distribution of (4.9), with ρ instead of ρ_1. In equation (4.9), t was defined as $n^{1/2}(z + \rho_1) / \chi$, but we can easily see that if the signs of both t and ρ_1 are changed, f(t) is unaltered, and we note that the double change brings us to the definition of (5.10).

If ρ is known, the value of u (say u_α) such that $\Pr\{u > u_\alpha\} = P$ is given by

$$u_\alpha = t_\alpha , \tag{5.11}$$

where t_α depends on n, ρ and P, and can be found from the tables of non-central t. Hence, if we find u for a sample of N items from the population, we can fix confidence limits for ρ by supposing that ρ is such that $u = u_\alpha$. We read from the tables the value of ρ corresponding to the given n, t(=u) and P, and our upper confidence limit for π (supposing that $P < 0.5$) is $1 - \Phi(N^{1/2} \rho)$.

The lower confidence limit is found similarly by replacing P by 1 - P. The confidence coefficient corresponding to these limits is 1 - 2P.

Suppose for example that we want 90% confidence limits for the proportion of defectives in the population, as determined from a sample of size 50. We will agree to regard an object as defective if the value of X for this object exceeds 1.645. From the sample we find, say, $\bar{x} = 0.14$, $s = 0.90$, so that $u = 7(1.505)/0.90 = 11.7$.

Putting $P = 0.05$, $t = 11.7$, $n = 49$, we find from the tables of Johnson and Welch that $\rho = 9.12$, so that $N^{-1/2} \rho = 1.29$ and the upper confidence limit for π is 0.099. Putting $P = 0.95$, we get $N^{-1/2} \rho = 2.01$, so that the lower confidence limit is .022. The estimate of π given by $1 - \Phi(N^{-1/2} u)$ is 0.050.

Instead of being given x_0, we may ask what value it should have in order to correspond to a given value of π. By our assumptions,

$$\pi = 1 - \Phi\left(\frac{x_0 - \mu}{\sigma}\right) = 1 - \Phi\left(\rho N^{-1/2}\right) , \tag{5.12}$$

so that $x_0 = \mu + \sigma z_\pi$, where z_π is the standard normal variate exceeded with probability π, and is equal to $\rho N^{-1/2}$. An estimate of x_0 is therefore given by

$$\hat{x}_0 = \bar{x} + s\left(\frac{N}{N-1}\right)^{1/2} z_\pi = \bar{x} + s n^{-1/2} \rho . \qquad (5.13)$$

In the above example, if $\pi = 0.05$, $z_\pi = 1.645$, and $\hat{x}_0 = 0.14 + 0.90\left(\frac{50}{49}\right)^{1/2}(1.645) = 1.635$.

We can again use the non-central t distribution to find confidence limits for x_0.

If $u > u_\alpha$, $n^{1/2}(x_0 - \bar{x}) > s u_\alpha$,

so that $x_0 > \bar{x} + s u_\alpha n^{-1/2}$.

If therefore we find t_α corresponding to the given values of n, ρ and P we can put $u_\alpha = t_\alpha$ and calculate x_0. By taking P = 0.05 and 0.95 we get upper and lower confidence limits for x_0.

Thus, in the same example as before, $\rho = 1.645\,(50)^{1/2} = 11.63$, n = 49 and P = 0.05. From the tables we can find $t_\alpha = 14.60$, so that $x_0 > 0.14 + \frac{0.90}{7} \times 14.60 = 2.02$. If P = 0.95, we get $t_\alpha = 9.40$, so that $x_0 > 1.21$.

The 90% confidence limits are therefore 1.21 and 2.02.

VI THE F-TEST

6.1 Comparison of Variance for two Normal Populations

Suppose we wish to compare the variances σ_1^2 and σ_2^2 for two populations, known to be normally distributed. Let the null hypothesis H_o be that $\sigma_1 = \sigma_2$ and the alternative hypothesis H_1 that $\sigma_1 = \lambda\sigma_2$ where we may take $\lambda > 1$. The usual test is to compute the function:

$$F = \frac{N_1 s_1^2}{N_2 s_2^2} \cdot \frac{N_2 - 1}{N_1 - 1} \quad , \tag{6.1}$$

which is a ratio of an unbiased estimate of σ_1^2, given by the sample variance s_1^2, to the corresponding estimate of σ_2^2, N_1 and N_2 being the sizes of the two samples.

The distribution of F on the null hypothesis is known. Its probability density is

$$f(F) = \frac{(n_1/n_2)^{\frac{n_1}{2}} F^{\frac{n_1-2}{2}}}{B(n_1/2, n_2/2)(1 + n_1 F/n_2)^{\frac{n_1+n_2}{2}}} \tag{6.2}$$

where $n_1 = N_1 - 1$ and $n_2 = N_2 - 1$.

The hypothesis H_o is rejected if $F > F_\alpha$, where

$$\int_{F_\alpha}^{\infty} f(F)\, dF = \alpha \tag{6.3}$$

The probability of rejecting H_o when it is really true is then equal to α.

The power of the test is given by

$$P = Pr\left\{ F > F_\alpha \mid H_1 \right\} \tag{6.4}$$

Now, on the hypothesis H_1, the ratio $\frac{F\sigma_2^2}{\sigma_1^2} = \frac{N_1 s_1^2}{n_1\sigma_1^2} \Big/ \frac{N_2 s_2^2}{n_2\sigma_2^2}$ has the F distribution, so that if F_P is the value of F which there is a probability P of exceeding,

$$Pr\left\{ \sigma_2^2 F / \sigma_1^2 > F_P \right\} = P,$$

$$\text{i.e. Pr}\left\{ F > \sigma_1^2 F_P / \sigma_2^2 \right\} = P. \tag{6.5}$$

Comparing (6.4) and (6.5) we see that

$$F_\alpha = \sigma_1^2 F_P / \sigma_2^2 = \lambda^2 F_P . \tag{6.6}$$

We can use the tables of the F distribution calculated by Merrington and Thompson (1943) which are available in abridged form in Hald's Statistical Tables and Formulas, 1952, in order to calculate λ Thus, if $N_1 = N_2 = 10$, $\alpha = .05$, and $P = 0.5$, we find $F_\alpha = 3.18$, $F_P = 1.00$, so that $\lambda = (3.18)^{1/2} = 1.78$.

This means that we stand an even chance of recognizing a difference between σ_1 and σ_2 when the actual ratio is 1.78, provided we agree to accept a 5% chance of wrongly rejecting the null hypothesis when actually $\sigma_1 = \sigma_2$. Only for a few values of P can F_P be obtained directly from the tables. When N_1 and N_2 are reasonably large (say 30 or more) an approximation devised by A. H. Carter (1947) may be used to obtain F for other values. This approximation is actually for Fisher's z, which is more nearly normal than F, but since $z = 1/2 \log_e F$, F is readily obtained from a table of natural logarithms. The approximation consists in finding τ, the normal variate corresponding to P, and calculating

$$z \approx \frac{\tau (h+k)^{1/2}}{h} - \left(\frac{1}{n_1} - \frac{1}{n_2}\right)\left(k + \frac{5}{6} - \frac{s}{3}\right) \tag{6.7}$$

where $s = \frac{1}{n_1} + \frac{1}{n_2}$, $h = \frac{2}{s}$, $k = (\tau^2 - 3)/6$.

For $n_1 = n_2 = 19$, and $P = 0.25$, we have $\tau = 0.6745$, $k = -0.4242$, $h = 19$, and therefore $z \approx 0.1530$. This is equivalent to $F = 1.3580$, whereas the correct value is 1.369. In Table VII the entries for $P = 0.4$ and 0.6 have been calculated by means of this approximation, and checked by a set of curves drawn by Ferris, Grubbs and Weaver (1946).

Table VII gives values of λ for different sized samples (the two samples are supposed equal in size, so that $N_1 = N_2 = N$), and for certain values of P. This table can be used to decide roughly what size of samples would be necessary in order to detect a given difference of the value of λ from unity.

TABLE VII

Values of the Standard Deviation Ratio $\lambda = \sigma_1 / \sigma_2$, detectable with power P, the samples being both of size N.

N \ P	0.1	0.25	0.4	0.5	0.6	0.75	0.9	0.95
5	1.25	1.76	2.24	2.53	2.86	3.63	5.12	6.39
10	1.14	1.41	1.64	1.78	1.94	2.25	2.78	3.18
15	1.11	1.31	1.47	1.58	1.68	1.89	2.24	2.48
20	1.09	1.26	1.39	1.47	1.56	1.72	1.99	2.17
25	1.08	1.22	1.34	1.41	1.48	1.62	1.84	1.98
30	1.07	1.20	1.30	1.36	1.43	1.55	1.74	1.86
61	1.05	1.13	1.20	1.24	1.28	1.35	1.46	1.53
101	1.04	1.10	1.15	1.18	1.21	1.26	1.34	1.39

The probability of error of the first kind is 0.05; that of error of the second kind is 1 - P.

Thus, suppose that a suggested new proces of manufacture is expected to be able to reduce the dispersion in some quantity measured (e.g. tensile strength) for certain types of casting. If the new process can reduce dispersion in the ratio 3 : 2, it will be worth while changing. In order to have a 75% chance of detecting such a difference, if it exists, we need samples of about 35. In order to have a 95% chance, we should need samples of about 70.

6.2 The Analysis of Variance Test

The standard analysis of variance test is an F test of the hypothesis that there is no significant difference between the "treatments" being compared. The test is based on a comparison of two independent estimates of variance, one calculated from the treatment effects and one from the "error".

To take a simple case, suppose we have b treatments, each replicated r times. If X is the variable measured, and if x_{ij} is the observed value for the i^{th} treatment and the j^{th} replicate, the total sum of squares is

$$Q = \sum_{i,j} (x_{ij} - \bar{x})^2 = q_1 + q_2 ,$$

where $q_1 = r \sum_{i} (\bar{x}_{i.} - \bar{x})^2$ and $q_2 = \sum_{i,j} (x_{ij} - \bar{x}_{i.})^2$

Here $\bar{x}_{i.} = \frac{1}{r} \sum_{j} x_{ij}$, the mean value for the i^{th} treatment, and $\bar{x}$ is the overall mean. The quantities q_1 and q_2 are called the sum of squares between treatments and the sum of squares within treatments respectively. The latter depends only on the minor variations between replicates undergoing the same treatment, and the corresponding mean square $q_2/(r - 1)b$ is the error estimate of variance. The former depends on the average treatment effects, and the mean square $q_1/(b - 1)$ is the estimate of variance based on treatment differences.

On the null hypothesis, $\dfrac{q_1}{b - 1} \cdot \dfrac{b(r - 1)}{q_2} = F$ has the F distribution with n_1 and n_2 degrees of freedom, where $n_1 = b - 1$ and $n_2 = b(r - 1)$. Since

$$F = n_2 q_1 / n_1 q_2, \qquad \frac{n_1 F}{n_1 F + n_2} = \frac{n_1 n_2 q_1}{n_1 n_2 (q_1 + q_2)} = q_1 / Q.$$

This quantity, denoted by E^2, has been used in the tables prepared by P. C. Tang (1938) for the power function of the analysis of variance test.

On the alternative hypothesis that the true effect of the i^{th} treatment is α_i (where we may suppose the origin so selected that $\sum \alpha_i = 0$), the variance of treatment effects is

$$\sigma_T^2 = \frac{1}{b} \sum_i \alpha_i^2 \tag{6.8}$$

If the true variance of the population sampled is σ^2 (irrespective of the magnitude of the treatment effects), the variance of a treatment mean is $\sigma_M^2 = \sigma^2/r$. The ratio

$$\phi = \sigma_T / \sigma_M \tag{6.9}$$

is used in Tang's tables as an argument. When $\phi = 0$ the quantity q_1/σ^2 has the ordinary χ^2 distribution with n_1 degrees of freedom. When $\phi \neq 0$, it has a <u>non-central</u> χ^2 distribution, the probability density being

$$f(\chi'^2) = \frac{1}{2} e^{-\frac{\lambda}{2}} \left(\frac{\chi'^2}{2}\right)^{\frac{b-3}{2}} e^{-\chi'^2/2} K(\lambda \chi'^2/4), \tag{6.10}$$

where χ'^2 is written for q_1/σ^2, and λ for $b\phi^2$, and where K(x) is an infinite series:

$$K(x) = \sum_{m=0}^{\infty} \frac{x^m}{m!\,\Gamma(m + (b-1)/2)} = \frac{1}{\Gamma\left(\frac{b-1}{2}\right)} \left[1 + \frac{2x}{1!\,(b-1)} + \frac{(2x)^2}{2!\,(b-1)(b+1)} + \cdots \right] \tag{6.11}$$

When $\lambda = 0$, $K(\lambda\chi'^2/4)$ in (6.10) reduces to $K(0) = \frac{1}{\Gamma\left(\frac{b-1}{2}\right)}$ and (6.10) is then the ordinary χ^2 density function with b - 1 degrees of freedom.

The density function for E^2, which is obtainable from the non-central χ^2 density for q_1/σ^2 and the central χ^2 density for q_2/σ^2, is given by

$$f(E^2) = \frac{e^{-\lambda/2}(E^2)^{\frac{1}{2}n_1 - 1}(1 - E^2)^{\frac{1}{2}n_2 - 1}}{B(\frac{1}{2}n_1, \frac{1}{2}n_2)} H(\lambda E^2/2) \qquad (6.12)$$

where

$$H(x) = 1 + \frac{n_1 + n_2}{n_1}\frac{x}{1!} + \frac{(n_1+n_2)(n_1+n_2-2)}{n_1(n_1+2)}\frac{x^2}{2!} + \cdots \qquad (6.13)$$

and where $n_1 = b - 1$, $n_2 = br - b$, $n_1 + n_2 = br - 1$. The function $H(x)$ is called a confluent hypergeometric function. When $\lambda = 0$, that is when $\phi = 0$, we have $H(\frac{1}{2}\lambda E^2) = 1$, and E^2 is a Beta-variate. The probability of error of the first kind, if the null hypothesis is rejected when $E^2 > E_\alpha^2$, is given by

$$\alpha = \int_{E_\alpha^2}^{1} f(E^2 \mid \lambda = 0)\, dE^2 \qquad (6.14)$$

For a given α, E_α^2 is found from the tables of the Incomplete Beta Function. The power of the test is given by

$$P = 1 - \beta = \int_{E_\alpha^2}^{1} f(E^2)\, dE^2 \qquad (6.15)$$

and can be calculated numerically, $f(E^2)$ being now given by (6.12), with $\lambda \neq 0$. Tang's tables give, for different numbers of degrees of freedom, the corresponding values of E_α^2 for $\alpha = 0.05$ and 0.01, and also the power P for selected values of ϕ. Emma Lehmer (1944) has published inverse tables which for selected values of P give the corresponding values of ϕ, both for $\alpha = 0.05$ and for $\alpha = 0.01$.

In the special case when $b = 2$ (and therefore $n_1 = 1$), there is only one set of α_i which will yield a given ϕ and at the same time satisfy the relation $\sum \alpha_i = 0$. If the true treatment means for the first and second treatments are $\mu_0 = \mu + \alpha_1$, and $\mu_1 = \mu + \alpha_2$, we clearly have $-\alpha_1 = \alpha_2 = 1/2(\mu_1 - \mu_0)$, so that $\sigma_\tau^2 = \frac{1}{4}(\mu_1 - \mu_0)^2$ and $\phi = \frac{\mu_1 - \mu_0}{2\sigma}\sqrt{r}$. The degrees of freedom n_1 and n_2 are 1 and $2(r - 1)$ respectively.

The quantity ρ of Neyman and Tokarska's tables (see § 4.5) is here $\rho = \frac{\mu_1 - \mu_0}{\sigma}\left(\frac{r}{2}\right)^{1/2}$, since $N_1 = N_2 = r$, so that $\rho = \sqrt{2}\,\phi$. However, Tang's tables are for the two-tailed test and Neyman and Tokarska's for the one-tailed test, so that Tang's (or Lehmer's) level $\alpha = 0.05$ corresponds to Neyman's level $\alpha = 0.025$.

6.3 An Example

This example is given by Tang. Suppose we have four treatments, each with five replications, in a randomized block experiment. Then $n_1 = 3$ and $n_2 = 16 - 4 = 12$ (four degrees of freedom are allowed between blocks). Let the treatment differences (expressed as percentages of the mean yield) be -5, -4, 3, 6, and let the standard deviation (estimated from past experience) be 10% of the mean. Then

$$\phi^2 = \left(\frac{1}{4}\sum \alpha_i^2\right)\left(5/\sigma^2\right)$$

$$= \frac{5}{4} \cdot \frac{86}{100} = 1.075, \text{ so that } \phi = 1.04.$$

Tang's tables for the 5% significance level show that when $\phi = 1$, $P = 0.269$ and when $\phi = 1.5$, $P = 0.556$. This suggests that P is about 0.3, so that if the true treatment differences were as indicated above, the chance of detecting them at the 5% level would be only about 3 out of 10.

A similar result holds for a Latin square experiment. In a square of size n x n, the degrees of freedom are $n_1 = n - 1$, $n_2 = (n - 1)(n - 2)$.

6.4 The F-test, on the Assumption that the Treatment Effects are not Constant but are Normally Distributed

We can in some cases suppose that the b lot means represent a sample from a normally distributed super-population of means with a standard deviation $\theta\sigma$, θ being a pure number. The sampling variance for a single lot mean of size r is σ^2/r, so that the total variance among the means is

$$\sigma^2/r + \theta^2\sigma^2 \quad = \quad \lambda^2\sigma^2/r \quad , \tag{6.16}$$

where $\lambda^2 = 1 + r\theta^2$. The null hypothesis H_o is that $\theta = 0$, and the alternative hypothesis H_1 is that $\theta > 0$.

On hypothesis H_o, $\dfrac{b(r-1)q_1}{(b-1)q_2}$ has the F distribution with $b - 1$ and $b(r - 1)$ degrees of freedom. On H_1, the quantity $\dfrac{b(r-1)q_1}{\lambda^2(b-1)q_2}$ has the F distribution with the same degrees of freedom.

The hypothesis H_o will be rejected when $F > F_\alpha$, F here standing for $\left[b(r-1)q_1\right] / \left[(b-1)q_2\right]$. If H_1 is true, $F > F_\alpha$ implies that $F/\lambda^2 > F_\alpha/\lambda^2$, and the probability of this is the power of the test. Therefore, since F/λ^2 has the F distribution,

$$P = \int_{F_\alpha/\lambda^2}^{\infty} f(F)dF \tag{6.17}$$

and this can be found from tables of the F function.

Table VIII has been calculated from Merrington & Thompson's tables (1943) of the F distribution. It gives for $\alpha = .05$ and $P = 0.5$, the value of θ corresponding to selected values of b and r. That is to say, it gives the standard deviation of lot means, as a fraction of the standard deviation of the population, which has an even chance of being detected at the 5% significance level, as a result of an analysis of variance based on b lots with r replicates in each lot. Thus, with 3 lots, one would need at least 5 items in each sample in order to stand an even chance of finding a significant component of variance between lot means, if actually this component were as large as the population variance ($\theta = 1$).

TABLE VIII

Ratio of Standard Deviation of Lot Means to Standard Deviation of Population, detectable with Power 0.5 at Significance Level 0.05, for b Treatments (or Lots) with r Replicates.

r \ b	2	3	4	5	6	7	8	9	10
2	3.66	2.22	1.73	1.48	1.32	1.21	1.13	1.06	1.01
3	2.09	1.37	1.11	0.98	0.89	0.83	0.78	0.74	0.71
4	1.63	1.08	0.89	0.79	0.72	0.65	0.63	0.60	0.58
5	1.39	0.93	0.77	0.65	0.62	0.58	0.55	0.52	0.50
6	1.23	0.82	0.68	0.61	0.55	0.52	0.49	0.47	0.45
7	1.12	0.75	0.62	0.55	0.51	0.47	0.45	0.43	0.41
8	1.04	0.69	0.58	0.51	0.47	0.44	0.41	0.40	0.38
9	0.97	0.65	0.54	0.48	0.44	0.41	0.39	0.37	0.36
10	0.91	0.61	0.51	0.45	0.41	0.39	0.37	0.35	0.34
11	0.86	0.58	0.48	0.43	0.39	0.37	0.35	0.33	0.32
12	0.82	0.55	0.46	0.41	0.37	0.35	0.33	0.32	0.30
16	0.70	0.47	0.39	0.35	0.32	0.30	0.28	0.27	0.26
21	0.61	0.41	0.34	0.30	0.28	0.26	0.25	0.24	0.23
25	0.56	0.37	0.31	0.28	0.25	0.24	0.22	0.21	0.21
31	0.50	0.33	0.28	0.25	0.23	0.21	0.20	0.19	0.19
61	0.35	0.24	0.20	0.17	0.16	0.15	0.14	0.14	0.13

VII. DISTRIBUTION-FREE TESTS

7.1 Van der Waerden's Test for the Difference of Means of two Samples

The chief objection to Student's test for the difference of means is the necessity for assuming normality in the distributions. Various tests have been derived which do not require this assumption, but such tests are usually considerably less powerful than Student's test, particularly when the samples are fairly large. Van der Waerden (1935) has described a test which does not require the assumption of normality but which, if the distributions are normal, is asymptotically as powerful as Student's test.

Suppose there are m observations $x_1, x_2, \ldots, x_m$ and n observations $y_1, y_2, \ldots, y_n$. Let the means of these two samples be $\bar{x}$ and $\bar{y}$ respectively and let $D = \bar{x} - \bar{y}$. The null hypothesis to be tested is that $D = 0$ and the alternative hypothesis is that $D > 0$.

The method consists in placing all the observations in order of increasing size (the x's and y's mixed up together), labelling them $z_1, z_2 \ldots z_N$ $(N = m + n)$, and associating with each z_k $(k = 1, 2, \ldots, N)$ a standardized normal variate ζ_k, defined by $\Phi(\zeta_k) = k/(N + 1)$. This means that the ζ_k are normal deviates corresponding to values of the cumulative distribution function which are evenly spaced between 0 and 1. Thus if $m = n = 5$, there are 10 values of ζ_k, given by $\Phi(\zeta_1) = 1/11$, $\Phi(\zeta_2) = 2/11 \ldots$, $\Phi(\zeta_{10}) = 10/11$. If we denote the inverse function by Ψ, we can write

$$\zeta_k = \Psi\left[k/(N + 1)\right] \tag{7.1}$$

where
$$(2\pi)^{-1/2} \int_{-\infty}^{\Psi(x)} e^{-z^2/2} = x \, . \tag{7.2}$$

We now pick out and total those values of ζ_k which are associated with the x's, and which we may denote by $\xi_1, \xi_2, \ldots, \xi_m$. If the x's on the whole are larger than the y's, the total $\Sigma \xi_i$ will be greater than zero. If this total exceeds a certain critical value, depending on m and n, the difference D between $\bar{x}$ and $\bar{y}$ may

be regarded as significantly different from zero. A table of critical values corresponding to the 5% significance level is given in Table IX.

Thus, suppose the following sample values are recorded:

x	26	29	28	24	22	
y	23	25	21	18	20	27

Here m = 5, n = 6, $D = \bar{x} - \bar{y} = 3.47$. The values are placed in order in the following table, with the x's ringed. The x_i are obtained conveniently from Kelley's Statistical Tables (1948).

z	18	20	21	(22)	23	(24)	25	(26)	27	(28)	(29)
k	1	2	3	4	5	6	7	8	9	10	11
$\xi_i = \Psi\left(\frac{k}{12}\right)$			-0.4307		0		0.4307		0.9674	1.3830	

The sum of the ξ_i is 2.3504, and the 5% critical value is 2.28. The difference between $\bar{x}$ and $\bar{y}$ is therefore significant at the 5% level.

This test depends only on the order of the observed values. There are actually 462 ways in which 5 x's and 6 y's can be permuted, counting as different only those ways in which the x's occupy different relative positions in the sequence. Of these ways, exactly 23 have $\sum \xi_i \geq 2.317$ and 24 have $\sum \xi_i \geq 2.278$. If all arrangements are equally probable, the chance of wrongly rejecting the null hypothesis is 23/462, which is very close to 0.05.

7.2 Calculation of Critical Values

For large values of N, the distribution of $\sum \xi_i$ approximates to normal with a mean of zero. The true variance is m n Q/(N - 1), where

$$Q = \frac{1}{N} \sum_{k=1}^{N} \left\{ \Psi\left(\frac{k}{N+1}\right) \right\}^2 \qquad (7.3)$$

TABLE IX

Critical Values of $\sum \xi_i$ for Van der Waerden's Test ($\alpha = 0.05$)

N \ (m_n or)	2	3	4	5	6	7	8	9	10	15	20
6	1.47	1.56									
7	1.56	1.70									
8	1.63	1.82	1.88								
9	1.68	1.91	2.01								
10	1.73	1.98	2.12	2.16							
11	1.77	2.04	2.21	2.28							
12	1.81	2.10	2.29	2.40	2.43						
13	1.83	2.14	2.34	2.47	2.53						
14	1.86	2.18	2.40	2.54	2.63	2.66					
15	1.86	2.21	2.45	2.61	2.71	2.76					
20	1.97	2.34	2.62	2.84	3.00	3.13	3.21	3.26	3.28		
25	2.02	2.42	2.73	2.98	3.18	3.35	3.48	3.58	3.65		
30	2.06	2.48	2.81	3.08	3.31	3.50	3.66	3.79	3.90	4.13	
35	2.09	2.52	2.87	3.15	3.40	3.60	3.78	3.94	4.07	4.46	
40	2.11	2.56	2.91	3.21	3.46	3.69	3.88	4.05	4.20	4.70	4.85

The size of the x sample is m and that of the y sample n (N = m + n). The x sample is that with the greater mean. If m > n, read the above table for n instead of m.

For small values of N this can easily be evaluated and for large N the approximation

$$\Omega \approx 1 - (2 \ell n N)/N + (\ell n \ell n N)/N \qquad (7.4)$$

(where ℓn is the Napierian logarithm) is remarkably good, as indicated in the following brief table:

N	Ω(exact)	Ω(approx.)
5	0.4486	0.4513
10	0.6216	0.6229
15	0.7045	0.7053
20	0.7546	0.7553

For the normal approximation, the critical value of $\sum \xi_i$ is taken as $\Psi (1 - \alpha) \left[m n \Omega / (N - 1) \right]^{1/2}$, where α is the probability of error of the first kind. In Table IX, $\alpha = 0.05$ and $\Psi (1 - \alpha) = 1.6449$.

The critical values in this table are calculated for the normal approximation.

For small values of N, the probability of error of the first kind, with these critical values, will not be exactly 0.05. However, the following table indicates that the differences from 0.05 are not serious as long as neither m nor n is very small.

TABLE X

Probabilities of Error of the First Kind in using the Critical Values of Table IX.

N \ m	2	3	4	5
6	.067	.050		
7	.048	.057		
8	.071	.054	.057	
9	.055	.048	.048	
10	.044	.050	.052	.048
11	.055	.055	.051	.050

7.3 The Power of the Van der Waerden Test

Since the test depends only on the order of the observations, the actual distribution function for the x's or the y's is irrelevant. If we suppose that the x's are normally distributed with mean a and variance 1 and the y's are also normally distributed with mean 0 and variance 1, then it can be shown that $\sum \xi_i$ has asymptotically (for fixed m and for $N \to \infty$) the same distribution as $\sum x_i$, and its standard deviation is $m^{1/2}$.

Asymptotically, Student's test consists in rejecting the null hypothesis (namely, that a = 0) when $m^{1/2} \bar{x} > \Psi(1 - \alpha)$, i.e. when $\sum x_i > m^{1/2} \Psi(1 - \alpha)$. The Van der Waerden test rejects H_o when $\sum \xi_i > \sigma \Psi(1-\alpha)$ where $\sigma^2 = m(N - m) Q/(N - 1)$, and since the distribution of $\sum \xi_i$ approaches that of $\sum x_i$ and σ^2 approaches m as $N \to \infty$, the two tests are evidently asymptotically equivalent.

7.4 Treatment of Ties

It may happen in practice that there are some ties in the ranking, because, even though the variates are continuous, the measurements are rounded off. These ties may be treated in different ways:

(1) if q values of z_k ($k = 1, 2, \ldots, N$) are equal to z_j and p of them are less than z_j, we assign at random the ranks $p + 1$, $p + 2$, ..., $p + q$ to the q equal values, and therefore the corresponding ζ's are taken as $\zeta_{p+1}, \ldots, \zeta_{p+q}$;

(2) we assign the ranks among the tied variates in all possible ways. If there are s of these ways and for r of them the value of $\sum \xi_i$ belongs to the critical region, we reject H_o with probability r/s;

(3) each set of q tied values may be alloted a ζ_j which is the arithmetic mean of the $\zeta_{p+1}, \ldots, \zeta_{p+q}$. This necessitates a reduction in the sum of squares of the ζ_k used in calculating Q, so that the critical value needs some adjustment. If, for example, the items ranked 8 and 9 are judged equal in a total of 11, they may be given the value $1/2(0.431 + 0.674) = 0.553$. The reduction in the sum of squares is $1/2(0.674 - 0.431)^2 = 0.0297$.

7.5 Terry's Test

M. E. Terry (1952) suggested independently a test which is very similar to Van der Waerden's. As applied to the two-sample problem, the null hypothesis H_o is that the two samples (of m and n observations) come from the same continuous population, and the test is most powerful against the alternative hypothesis H_1 that they come from two normal populations with means μ_1 and μ_2 and common variance σ^2, the ratio $(\mu_1 - \mu_2)/\sigma$ being sufficiently small.

The test consists in computing a statistic c, which is the sum of the expected values of those m items (in a sample of m + n drawn from a standard normal population) which have ranks the same as those of the x_i in the observed combined sample, when the x's and

y's are placed in order. Thus, in the illustration used in § 7.1, the x's occupy ranks 4, 6, 8, 10 and 11. From Table XX of Fisher and Yates's Statistical Tables, the expected values for these ranks in a sample of 11 from a standard normal population would be -0.46, 0, 0.46, 1.06, and 1.59 respectively, and the statistic c is therefore 2.65. A 5% critical value for c is determined by computing c for the 23 or 24 permutations out of 462 which give the largest values. It is found that 23 permutations have $c \geq 2.54$ and 24 have $c \geq 2.47$. The 5% critical value is therefore about 2.54, and accordingly the difference between the samples is significant at the 5% level.

Terry has shown that his statistic c has on the null hypothesis a variance $V(c) = \frac{mn}{N(N-1)} \sum \mu_i^2$, $N = m + n$, where μ_i is the expected value of the item of rank i in a sample of N drawn from a standard normal population. Values of $\sum \mu_i^2$ are given in Fisher and Yates's Table XXI. He has also shown that if $0 < n/N < 1$ as $N \to \infty$, then the distribution of c approaches normal with mean 0 and variance V(c).

The distribution of $c(N-2) / \left[V(c)(N-1) - c^2 \right]$ is approximately that of Student's t, with N - 2 degrees of freedom. Thus, for the case given above, with m = 5, n = 6, and N = 11, the value of $t_{0.05}$ for 9 d.f. is 1.833 (the one-tailed test), and $\sum \mu_i^2 = 8.8892$. The critical value of c_1, determined by

$$c_1^2 \left(N - 2 + t_{0.05}^2 \right) = \frac{mn}{N} \sum \mu_i^2 \, t_{0.05}^2 \qquad (7.5)$$

is 2.57, which agrees very well with the correct value 2.54. The power of this test has not been determined analytically, but experimental results on random numbers indicate that the power is not far below that of Student's t, even for N as small as 8, when $(\mu_1 - \mu_2) / \sigma$ is less than 0.5 or greater than 2.5. For intermediate values there is a marked reduction in power.

Ties are treated as in Van der Waerden's test.

7.6 The Mann and Whitney (or Wilcoxon) Test

This is a rank order test of the hypothesis H_o that two sets of sample values $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ come from the same population, against the alternative hypothesis H_1 that the x's

are stochastically larger than the y's. If the random variables X and Y have continuous cumulative distribution functions $F(x)$ and $G(y)$, X is stochastically larger than Y when for every a, $F(a) < G(a)$, i.e. the probability that $X \leq a$ is smaller than the probability that $Y \leq a$.

A statistic T to test this was first proposed by Wilcoxon (1945). If the x's and y's are arranged in increasing order, T is the sum of the ranks of the x's in this sequence. For the data given in § 7.1, these ranks are 4, 6, 8, 10, 11, so that T = 39. An equivalent statistic U was tabulated by Mann and Whitney (1947) and is the number of inversions, that is, the number of times an x precedes a y. In the example, U = 3 + 2 + 1 = 6, since the value x = 22 precedes the three values y = 23, 25 and 27, x = 24 precedes y = 25 and 27, and so on. In general

$$U = mn + m(m + 1)/2 - T. \tag{7.6}$$

If, under the null hypothesis, $\Pr(U \leq \bar{U}) = \alpha$, the test which consists in rejecting H_o when $U \leq \bar{U}$ has a size α.

The expectation of U on the null hypothesis is $nm/2$ and its variance is $nm(n + m + 1)/12$. The limiting distribution of U is normal as both m and n tend to ∞, and for $m = n = 8$ the distribution of $U - 1/2\, n\, m$ is very close to normal. Mann and Whitney have calculated tables, for m and n not greater than 8, giving the probabilities of obtaining different possible values of U. Thus, for m = 5 and n = 6 (m and n can be interchanged in the tables) we find that $\Pr(U \leq 5)$ is 0.041 and $\Pr(U \leq 6)$ is 0.063. If we reject H_o when $U \leq 6$ (and therefore will do so in the example given in § 7.1) the probability of error of the first kind is 0.063. If we want to keep this error below 0.05 we shall have to take U = 5, and the hypothesis H_o will, in this particular example, not be rejected. Alternatively, we could reject it with probability 0.41 (since 0.041 + 0.41 (0.063 - 0.041) = 0.05), using a table of random numbers.

The maximum values of $\bar{U}$ such that the size of the test is not greater than 0.05 are given in Table XI. For larger m and n the normal approximation may be used

$$\bar{U} = 1/2(nm - 1) - 1.645 \left[\frac{nm(n + m + 1)}{12}\right]^{\frac{1}{2}} \qquad (7.7)$$

(The term -1/2 arises because of the discontinuity of U).

For n = m = 8, this give $\bar{U} = 15.8$. The actual probability is 0.052 for a value less than or equal to 16 and 0.041 for a value less than or equal to 15.

TABLE XI

Critical Values $\bar{U}$ for $\alpha \leq 0.05$, for given sample sizes m and n ($m \leq n$), in the Mann and Whitney Test.

n \ m	2	3	4	5	6	7	8
3	0	0					
4	0	0	1				
5	0	1	2	4			
6	0	2	3	5	7		
7	0	2	4	6	8	11	
8	1	3	5	8	10	13	15

The Mann and Whitney test is less powerful than the Van der Waerden or the Terry test. The test is consistent, in the same sense that as m and n tend to infinity the probability of rejection of the null hypothesis, when the alternative hypothesis is true, tends to 1.

7.7 The Sign Test

This is an approximate test of the difference between two sets of paired observations. The usual test, assuming normality of the distributions, is the Student t-test for the mean of the differences between pairs, the null hypothesis being that this mean is zero. The sign test takes account merely of the signs of these differences. The two observations belonging to one pair are assumed made under conditions as alike as possible, but conditions may vary widely from one pair to another, and this circumstance may invalidate the Student t-test.

It is assumed that the + and - signs of the paired differences have, on the null hypothesis, a binomial distribution. Zero differences are ignored. If there are N non-zero differences, with x of these positive and N - x negative, the hypothesis H_o is that in independent sampling x is distributed according to the terms of the binomial $(1/2 + 1/2)^N$. The alternative hypothesis H_1 is that the distribution of x is according to $(q + p)^N$, where $p \neq 1/2$. [1] If r is the smaller of x and N - x the test consists in rejecting H_o when $r \leqslant r_\alpha$, r_α being a number which depends on N and on the assumed significance level α.

A table of critical values for the application of the test was compiled by Dixon and Mood (1946) and is reproduced in Dixon and Massey's "Introduction to Statistical Analysis" (McGraw Hill, 1951). A discussion of the power of the sign test was given by W. M. Stewart (1941).

The following questions arise: (a) what is the minimum value of N that must be used if we want a given power for testing H_o against some assumed alternative H_1? (b) what is the maximum value that r may have for a given N if H_o is to be rejected at significance level α? Table XII is extracted from Stewart's paper. It shows, for example, that in order to have at least an even chance of detecting, at the 5% significance level, the difference between p = 0.70 and p = 0.50, one would need at least 25 pairs, and the lesser number of like signs must not be more than 7. Suppose, for example, machined parts of a specified diameter are tested by a "go" and "no-go" gauge, and on the average 50% will "go". A new machine produces parts of which 70% go. To have an even chance of finding a significant difference at the 5% level, one would need a sample of at least

25, and, to be significant, 18 or more should "go".

TABLE XII

Minimum N and Maximum r for testing with Power P the Hypothesis that the Proportion of Signs in Paired Differences is 0.5, the True Value of p being as given.

P \ p	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
.30	56, 20	25, 7	18, 4	9, 1	11, 1	7, 0		
.50	101, 40	44, 15	25, 7	18, 4	13, 2	10, 1	6, 0	
.70	158, 66	67, 25	40, 13	25, 7	18, 4	12, 2	10, 1	6, 0
.95	327, 145	143, 59	79, 30	49, 17	35, 11	23, 6	17, 4	12, 2

In each pair N is the first number and r the second. For $p < 0.50$ use $1 - p$. If x is the number of + signs and $N - x$ the number of - signs, r is the smaller of x and $N - x$.

REFERENCES

Carter, A.H., "Approximation to Percentage Points of the z Distribution", Biometrika, 34, 1947, pp. 352-358.

Dantzig, G.B., "The Non-existence of a Test of Student's Hypothesis having a Power Function independent of σ ", Ann. Math. Stat., 11, 1940, pp. 186-192.

Dixon, W.J. and Mood, A.M., "The Statistical Sign Test", Journ. Amer. Stat. Assoc., 41, 1946, pp. 557-566.

Ferris, C.D., Grubbs, F.E. and Weaver, E.L. "Operating Characteristics for the Common Statistical Tests of Significance", Ann. Math. Stat., 17, 1946, pp. 178-197.

Fisher, R.A. and Yates, F., Statistical Tables for Biological, Agricultural and Medical Research, (Oliver and Boyd, 3rd ed., 1949).

Fix, E., Tables of Non-central χ^2, (University of California Press, 1949).

Hald, A., Statistical Tables and Formulas, (Wiley, 1952)

Johnson, N.L. and Welch, B.L., "Applications of the Non-central t-distribution", Biometrika, 31, 1939, pp. 362-389.

Kelley, T., Statistical Tables (Harvard University Press, 1948).

Kenney, J and Keeping, E.S., Mathematics of Statistics, Part II (Van Nostrand, 1951).

Lehmann, E.L. and Stein, C., "Most Powerful Tests of Composite Hypotheses I. Normal Distributions", Ann. Math. Stat., 19, 1948, pp. 495 - 516.

Lehmer, E., "Inverse Tables of Probabilities of Errors of the Second Kind", Ann. Math. Stat., 15, 1944, pp. 388-398.

Mann, H.B. and Whitney, D.R., "On a Test of Whether One of two Random Variables is Stochastically larger than the Other", Ann. Math. Stat. 18, pp. 50-60.

Merrington, M., and Thompson, C.M., "Tables of Percentage Points Of the Inverted Beta (F) Distribution", Biometrika, 33, 1943, pp. 73-88.

National Bureau of Standards, Tables of the Binomial Probability Distribution, (Washington, D.C., 1950)

National Bureau of Standards, Tables of Probability Functions, vol. II (New York, Federal Works Agency, 1942).

Neyman, J. and Pearson, E.S., "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference", Biometrika, 20A, 1928, pp. 175-240 and 263-294.

Neyman, J. and Pearson, E.S., "On the Problem of the Most Efficient Tests of Statistical Hypotheses", Phil. Trans. Roy. Soc. A, 231, 1933, pp. 289-337.

Neyman, J. and Tokarska, B., "Errors of the Second Kind in Testing Student's Hypothesis", Journ. Amer. Stat. Assoc., 31, 1936, pp. 318-326.

Pearson, K., Tables of the Incomplete Gamma Function (H.M. Stationery Office, London, 1922).

Romig, H.G., "50-100 Binomial Tables" (Wiley, 1953).

Stewart, W.M., "A note on the power of the Sign Test", Ann. Math. Stat., 12, 1941, pp. 236-239.

Tang, P.C., "The Power Function of the Analysis of Variance Tests, with Tables and Illustrations of their Use", Statistical Research Memoirs (University College, London), 2, 1938, pp. 126-137.

Terry, M. E., "Some Rank Order Tests which are Most Powerful against Specific Parametric Alternatives", Ann. Math. Stat., 23, 1952, pp. 346-366.

Thompson, C. M., "Tables of the Chi-Square Distribution", Biometrika, 32, 1941-42, pp. 187-191.

Van der Waerden, B. L., "Ein neuer Test für das Problem der zwei Stichproben", Math. Annalen, 126, 1953, pp. 93-107.

Wald, A., "An Essentially Complete Class of Admissible Decision Functions, Ann. Math. Stat., 1947, pp. 549-555.

Wilcoxon, F., "Individual Comparisons by Ranking Methods", Biometrics Bulletin, 1, 1945, pp. 80-83

Wilks, S., "The Large Sample Distribution of the Likelihood Ratio for testing Composite Hypotheses", Ann. Math. Stat., 9, 1938, pp. 60-62.

Wilson, E. B. and Hilferty, M. M., "The distribution of Chi-square", Proc. Nat. Acad. Sci., 17, 1931, pp. 684-688.